Characteristics and quality of physical activity apps which provide feedback on user affect

A systematic review and evaluation of public and academic apps

Laura LAMMING

Submitted for the Degree of Master of Philosophy

Faculty of Health Studies

University of Bradford

2019

Abstract

Laura Lamming

Characteristics and quality of physical activity apps which provide feedback on user affect

A systematic review and evaluation of public and academic apps

Keywords: Physical Activity, Exercise, Behaviour Change, App, Smartphone, Mood, Affect, Emotion, Quality

Despite its benefits to both physical and mental health, physical activity levels worldwide remain low and new solutions for behaviour change must be sought. Smartphone apps are extremely popular and prevalent across the population, however their quality is still questionable. Physical activity produces an acute ‘feel good’ effect and intervention designers should consider the role that affect (mood) plays in uptake and maintenance of behaviours. It is timely to examine the use of affect as a motivator for physical activity, using new tools that allow real-time capture of both affect and physical activity (smartphones). The existence, characteristics and quality of physical activity apps that provide feedback on affect were explored in this thesis.

A mixed methods approach, comprising a systematic review (study 1) and a systematic evaluation (study 2) was taken. Data collection methods included both quantitative and qualitative assessments, using pre-existing and fit-for-purpose tools.

Twenty-two physical activity apps that provided feedback on affect were identified. Apps often failed to target groups most at risk of poor physical activity levels. Feedback on affect was performed in a variety of ways.

Quality of apps, based on 13 criteria, was mixed. Recommendations are made for researchers, app developers and funders, including the need for development of high quality physical activity apps incorporating and emphasising affective benefits, consideration of archiving processes for developed apps once development ceases, and collaboration between researchers, developers and users when designing apps.

Acknowledgements

With special thanks to Ian Kellar and Mohammed A. Mohammed, without your support I wouldn’t have started. Enormous hugs and thanks to Maria Horne, Neil Small, Andy Scally, Jane Montague, Muhammed Faisal, Katy Shire, Angela Richardson, Wendy Andrusjak, Jamie Beck, Sarah Farnell, Yvonne Messenger and Dan Mason, without whom I wouldn’t have finished.

Thank you to my ‘book club’ Katy, Bella and Noortje who taught me how to be a good postgrad. And to Dan, James and Ian who originally showed me how fun working in research could be. Thanks to my fellow postgrads for the camaraderie.

With thanks to my partner James (PhD) who inexplicably started dating me mid-thesis, just after he’d completed his own and therefore knew exactly what he was about to witness. Thank you for your patience, empathy and for cheering me on.

Thank you to my friends, in particular Sara, Cath, Ben, Emily, Lorna, the Team and the Guild who reminded me about life beyond my laptop.

Finally, great love and thanks to my family who were so far in my corner I couldn’t get in there myself. You’re the best.

Table of contents

List of Figures

List of Tables

Glossary of terms

Affect - short-lived feelings, both positive and negative.

Literature-based apps – apps that were identified within the literature-base, most likely developed primarily by academics with some support from app developers.

Public apps – apps that were identified from the public app stores including Google Play and the Apple App store. Most likely developed by commercial companies.

List of abbreviations

Abbreviation	Term
2D	Two dimensional
ABC	Antecedent, Behavioural, Consequences
API	Application Program Interface
BAME	Black, Asian and Minority Ethnic
BCT	Behaviour Change Technique
BCTTv1	Behaviour Change Technique Taxonomy version 1
BMI	Body Mass Index
CBT	Cognitive Behavioural Therapy
DP	Data Protection
DPA	Data Protection Act
ECG	Electrocardiogram
GDPR	General Data Protection Regulation
GP	General Practitioner
GPS	Global Positioning Service
HR	Heart Rate
HTTPS	Hypertext Transfer Protocol Secure
M	Mean
MARS	Mobile App Rating Scale
MoA	Mechanism of Action
MVPA	Moderate to Vigorous Physical Activity
NA	Non-applicable
NDSR	Nutrition Data System for Research
NHS	National Health Service
ODL	Observations of Daily Living
OSM	Open Street Map
PA	Physical Activity/Physical activity targeted by app
PA+	Physical activity and other behaviours targeted by app
PES	Player Experience Scale
PHCP	Primary Health Care Provider
PMS	Pre-menstrual Stress
PPI	Patient and Public Involvement
RCT	Randomised Control Trial
RMSE	Root Mean Square Error
SCT	Social Cognitive Theory
SD	Standard Deviation
SMS	Short Messaging Service
SUS	System Usability Scale
TCS	Theory Coding Scheme
TTFF	Time To First Fix
VPA	Vigorous Physical Activity
WA	Western Australia
WHO	World Health Organisation

1. Background

"*Psychology and humanity can progress without considering emotion—about as fast as someone running on one leg.*" (Russell, 2003, p. 145)

This chapter will provide an overview of physical activity and why it should be addressed. It will also describe evidence from previous attempts to change physical activity levels and propose that apps might be a good alternative approach. In particular, it will propose why using affect to do so could be effective and beneficial, reporting both evidence and theories to that effect. Finally, it will summarise the evidence for physical activity apps in general, current evidence for apps that provide feedback on affect and describe the research questions and approach of this thesis.

1.1 What is physical activity?

Physical activity (PA) is a complex behaviour. Before addressing the background and rationale for the following programme of work, the nature and definition of PA will briefly be described.

1.1.1 What is physical activity and why not exercise?

The terms "physical activity" and "exercise" have often been used interchangeably, but efforts have been made to distinguish between them (Caspersen et al., 1985). Exercise typically refers to scheduled, repetitive, purposeful activity, either for pleasure, or for the sake of becoming or staying fit and healthy (e.g. going to the gym, going for a run, playing a sport) (e.g. Caspersen et al., 1985). In contrast, while PA includes exercise (Caspersen et al., 1985), it also includes everyday activities such as walking to work or cleaning the house (World Health Organisation, 2010, 2018). Physical activity is the focus of this thesis.

Formally, physical activity has been defined as: *"any bodily movement produced by skeletal muscles that results in energy expenditure"* and can include *“sports, conditioning exercises, household tasks…and other activities”* (Caspersen et al., 1985, p. 127-128). Physical activity rather than exercise has been chosen for the focus of this thesis due to the range of acknowledged potential barriers associated with more structured exercise. Barriers include cost and being able to easily access equipment or services based on their proximity (Dishman et al., 1984; Sallis et al., 1997; Schutzer and Graves, 2004; Trost et al., 2002). Cultural barriers related to mixed-sex facilities (e.g. Lawton et al., 2006), and evidence that characteristics of activities can be less appealing or effective, are also mentioned in the literature (Dishman and Buckworth, 1996). For example, the ability to select your preferred activity and/or perform at a lower intensity level is associated with a more positive experience or with more activity (Dishman and Buckworth, 1996; Martinez et al., 2015; Parfitt and Gledhill, 2004; Trost et al., 2002; Zenko et al., 2016), which more structured exercise can restrict - the ‘rigidity’ of exercise programmes can be seen as a disincentive (Schneider et al., 2003). Survey data suggests walking and unstructured physical activities are popular compared to structured activities (Salmon et al., 2003). While exercise is often perceived as being a healthy option, general physical activities such as housework and gardening also contribute to promoting health and *“expend a significant amount of energy*” through non-exercise activity thermogenesis (NEAT) (Fujiki et al., 2008, p. 2). Physical activity encompasses many additional activities. A recent Health Survey for England reported the prevalence of such activities being carried out by adults, with 46% of respondents performing PA consisting of housework, 25% manual work, gardening, or DIY, and 47% walking (NHS Digital, 2017).

1.1.2 How much physical activity is enough?

The World Health Organisation (World Health Organisation, 2010) and UK Chief Medical Officer (Department of Health and Social Care, 2019) provide recommendations for PA for a range of age groups. However, as the current thesis will focus on adults, those are the guidelines that will be reported and typically referred to throughout this thesis unless otherwise specified (see figure 1 for recommendation and definitions). Amount, frequency and intensity of activity are characterised, all of which are important to incur the health benefits of physical activity.

Figure 1 UK Chief Medical Officer's recommendations for adults aged 19-64

Physical activity recommendations for adults aged 19–64 years old

1. For good physical and mental health, adults should aim to be physically active every day. Any activity is better than none, and more is better still.
2. Adults should do activities to develop or maintain strength in the major muscle groups. These could include heavy gardening, carrying heavy shopping, or resistance exercise. Muscle strengthening activities should be done on at least two days a week, but any strengthening activity is better than none
3. Each week, adults should accumulate at least 150 minutes (2 1/2 hours) of moderate intensity activity (such as brisk walking or cycling); or 75 minutes of vigorous intensity activity (such as running); or even shorter durations of very vigorous intensity activity (such as sprinting or stair climbing); or a combination of moderate, vigorous and very vigorous intensity activity.
4. Adults should aim to minimise the amount of time spent being sedentary, and when physically possible should break up long periods of inactivity with at least light physical activity.(Department of Health and Social Care, 2019)

Moderate-intensity physical activity: requires a moderate amount of effort and noticeably accelerates the heart rate.

Vigorous-intensity physical activity: requires a large amount of effort and causes a rapid breathing and a substantial increase in heart rate. (World Health Organisation, 2010)

1.2 Why look at physical activity?

1.2.1 We are not physically active

Physical activity is associated with numerous benefits including promotion of good physical health and prevention of non-communicable diseases (Lee et al., 2012), good mental health by both maintaining it (Kim et al., 2012) and facilitating recovery or improvements from conditions such as depression (Byrne and Byrne, 1993; Cooney et al., 2013; Schuch et al., 2016) and good quality of life (Anokye et al., 2012; Gill et al., 2013; Pucci et al., 2012). Despite these benefits, PA levels still remain low, not only nationally (NHS

Digital, 2017), but internationally (Hallal et al., 2012). To illustrate, a recent Health Survey for England (HSE) reported that only 66% of men and 58% of women met the guidelines for being aerobically active, and only 31% of men and 23% of women were achieving the recommended levels of muscle-strengthening activities (NHS Digital, 2017). These data were collected by questionnaire and therefore are likely to overestimate the levels of activity being achieved, thus underestimating the problem. For example, a previous HSE objectively monitored a subsample of those who completed a questionnaire and found that while 39% and 29% of men and women respectively met the recommendations for PA according to self-report, only 6% of men and 4% of women were achieving the recommendations when measured using an activity monitor (Joint Health Surveys Unit, 2009).

1.2.2 Population differences in physical activity levels: some people are more at risk than others of poor physical activity levels

While the problem of inactivity is international, there are variations within the world's population. These variations occur both within countries, for example in the UK, more adults tend to be active in London (65%), compared to the West Midlands (53%) (NHS Digital, 2017) as well as between countries, for example inactivity is higher in the Americas than in Southeast Asia (Hallal et al., 2012). There are also general variations between age groups and between genders, with activity decreasing with age and males being more active than females (Cooper et al., 2015; Hallal et al., 2012; Trost et al., 2002). Within populations, ethnic and socio-economic characteristics are also associated with variations in activity levels. For example, UK South Asians are less active than their White counterparts (Williams et al., 2011) with those with lower incomes typically being less active (Hallal et al., 2012; Wardle and Steptoe, 2003) and accumulating minutes of activity through everyday activities and walking rather than leisure-time activity (Ford et al., 1991). Although activity levels are of concern across the age, ethnicity and socio-economic spectrum, given the evidence presented, it could be argued that efforts should be targeted to adults of lower socio-economic status from non-white ethnicities.

1.2.3 There are health and economic costs of poor levels of physical activity

Being inactive has been reported as causing 6-10% of the major non-communicable diseases such as coronary heart disease, type 2 diabetes and some cancers, has reduced life expectancy and is as dangerous a risk factor for life expectancy as smoking or obesity (Lee et al., 2012).

Being inactive has poor outcomes both for individuals as well as for the population and the economy as a whole. In one of the first reports on the cost of global inactivity, Ding et al., (2016) conservatively estimated that it cost international health-care systems $53.8 billion in 2013, and in the UK, $1.8 billion. When indirect costs related to productivity losses from inactivity related deaths were included, the overall societal cost increased to $67.5 billion and $2.4 billion respectively, in the UK (Ding et al., 2016). Not only do these healthcare costs represent costs to the public sector ($1.5 billion) but also the private sector ($0.13 billion) and individual households ($0.17 billion) (Ding et al., 2016). It's clear that increasing population level activity would be of financial benefit to both the individual as well as health services.

1.3 Why are physical activity levels so low?

1.3.1 Modifiable and unmodifiable influences impact physical activity levels

Given the benefits, we must consider why physical activity levels are low. Barriers to PA have been explored extensively within the literature and are highly varied. A comprehensive discussion is beyond the scope of this thesis introduction. However, the barriers can be broadly divided into two categories, those that are modifiable and those that are not. Unmodifiable characteristics are those such as age or gender, typically genetically programmed characteristics, occupation, having children, marital status, race/ethnicity and many more (Doherty et al., 2017; NHS Digital, 2017; Trost et al., 2002). Researchers are not in a position to influence these characteristics ethically or practically, but they can be taken into consideration when developing and tailoring interventions.

In comparison, modifiable characteristics are those that can be targeted by PA interventions. They have been quite comprehensively collated in two reviews of correlates of PA, Trost et al., (2002) and Bauman et al., (2012) (see papers for full list of characteristics) and one review of sedentary behaviour that uses the socio-ecological approach to classify factors (O'Donoghue et al., 2016). They can include individual factors such as psychological, cognitive or emotional factors (e.g. attitudes, expected benefits, stress, self-motivation), or behavioural factors such as abilities or skills, habits, past and current behaviours. Interpersonal characteristics such as social and cultural factors, or features of the physical environment can also be included as modifiable factors. Bauman et al., (2012) also cite regional or national policies and global characteristics that could impact PA levels and could be classed as modifiable, for example transport systems and the global media or other megatrends such as urbanisation. Finally, the type and intensity of the PA is another obviously modifiable characteristic.

1.3.2 Influencing factors are complex

The previous section briefly outlines the range of characteristics that could influence PA levels, but neglects to capture the complex interplay of multiple variables and potential for moderating and mediating variables. In addition, it's likely that more characteristics and barriers will be found. For example, the advent of the industrial revolution drastically reduced the amount of manual labour required, motorised cars reduced travel time activity and the current phase of the technological revolution promises further reductions in occupational, domestic and travel-based activity, (Knuth and Hallal, 2009; Ng and Popkin, 2012), while sedentary behaviour is likely to increase through high screen time (Duncan et al., 2012). However, it could also be argued that the current climate crisis may reduce motorised travel.

Using the COM-B model (Michie et al., 2011, 2015), these characteristics can be summarised into three areas: characteristics related to an individual's capability, opportunity and motivation to perform the behaviour – physical activity. With such a range of potential barriers, it starts to become clear why optimal levels of PA are not more widespread.

1.4 What has been done to solve the problem so far?

Given the range of potential barriers to PA, an equally varied range of interventions have been created and tested. These can be broadly divided into population, community or group and individual level approaches (NICE, 2007). However, it's important to remember that such approaches can have an impact at different and multiple levels.

Population level approaches are those that seek change on a large scale, across many people, and measure that change at the level of the population. They can include wide-reaching campaigns and involve tools such as television, radio, print or social media, community activities, websites, school or worksite components and policy changes among other things (Noar, 2006). For example, This Girl Can is a UK based national initiative developed in 2015 to encourage girls and women to take part in physical activity (Sport England, 2015). Group-based interventions - *"social or family groups linked by networks, geographical location or another common factor"* (NICE, 2007, p. 33) - typically involve a number of participants receiving the intervention together, such as exercise classes, with outcomes at the group level. Perhaps the most commonly implemented by health researchers, is the individual level approach. These approaches directly reach out or provide an intervention to individual people and look for changes in people's behaviour/cognitions and so on. They can be delivered in a group-based context, but unlike group-based approaches, they look for changes within individuals.

Evidence for the effectiveness of each approach exists, but tends to be moderate at best, either for the whole intervention, or independent components, modes of delivery and for different populations (Cavill and Bauman, 2004; Cleland et al., 2012; Goto et al., 2014; Samdal et al., 2017; Wakefield et al., 2010; Webb et al., 2010). Each also has its limitations, such as the possibility for negative unintended consequences from large-reaching campaigns – such as desensitisation to health messages, and 'boomerang' effects where avoidance strategies start to appear (Cho and Salmon, 2007). They can also offer a 'one size fits all' approach which may not be appropriate to all participants. Evidence for cost-effectiveness is also questionable and group-based interventions may include dubious adaptations of typically individually-targeted content or components (Hoddinott et al., 2010).

Although there appears to be a dearth of literature that compares these three approaches, making it difficult to suggest one over the other, the socio-ecological model suggests that factors at the individual, community, and population level (intrapersonal, interpersonal, organisational or environmental, cultural/societal norms, laws and policies) all need targeting and should not be considered in isolation (McLeroy et al., 1988). Unfortunately, such a wide-reaching approach must be left to those with the resources and influence to enact it, which is beyond the scope of this thesis.

1.5 Why hasn't the problem been solved yet?

Despite successful outcomes at times, low PA is still a problem (Ekkekakis, 2017) and as recently as 2018, the World Health Organisation launched a programme targeting inactivity at all levels, via policy changes across countries using a systems-based approach (World Health Organisation, 2018). It could be argued that this programme takes a socio-ecological approach – targeting not just one level of intervention but all (McLeroy et al., 1988). The breadth and depth of this programme suggests why some of the previous approaches to inactivity reduction have not solved the problem. In particular, the need for wide reach but small changes, evidence-based practice and engaged and knowledgeable communities and individuals. The RE-AIM framework, developed to assess the public health impact of interventions also helps identify potential problems in past approaches, namely issues of reach, efficacy, adoption, implementation and maintenance (Glasgow et al., 1999).

1.5.1 The problem of reach

Previous attempts to change behaviour may not have been at a sufficiently wide-reaching level. Reach could become a problem, even if the intervention itself is effective, as the real-world introduces potentially unforeseen obstacles. Not only could the intervention fail to reach individuals in general, but it could also fail to reach high risk individuals or those most in need, increasing health inequalities (White et al., 2018).

1.5.2 The problem of effective content

Despite recent debates in the literature and mixed evidence, efficacy of an intervention is still thought to be related to using evidence-based and/or theory-based techniques and approaches (McEwan et al., 2018; Prestwich et al., 2015). Although this is becoming far more common, especially with the advent of tools such as the behaviour change taxonomy (Michie et al., 2013), it still remains a problem in areas of behaviour change – for example, many available apps in the app stores have limited theory-based content in them (Cowan et al., 2013). Nevertheless, large numbers of the population are downloading and using these apps (see 1.6.1).

1.5.3 The problems of implementation, adoption and maintenance

Availability of an intervention does not guarantee adoption. Individuals may find it unacceptable due to cost, mode of delivery, access to resources or dislike of the person delivering them. Alternatively implementation may be poor, incorrect or incomplete and intervention fidelity is an evolving area of focus in research (Moore et al., 2015). Finally, maintenance of the changed behaviour can be challenging. Many interventions focus on uptake, but this process is now considered different to that which maintains a behaviour (Kwasnicka et al., 2016; Stralen, 2011). Therefore, interventions may not be equipped to promote maintenance, only uptake, resulting in only a short-term change in activity levels with none of the meaningful longer lasting effects.

1.6 A possible solution: the rise of mHealth

One potential answer to some of the aforementioned issues regarding reach, efficacy, adoption, implementation and maintenance is the smartphone app.

1.6.1 Smartphones have reach as a mode of delivery

An intervention delivered by a smartphone has the potential to reach a lot of people, of varying characteristics, and integrates more easily into a user's daily life than a computer-based intervention (Gasser et al., 2006). Global smartphone ownership has been steadily rising since they first hit the market, with an estimated 2.8 billion users by 2020 (Statista, n.d.), while 71% of UK adults claimed to own a smartphone as of 2015 (Ofcom, 2015), and more recent studies suggesting this has risen by at least another 10% (Deloitte, 2016).

Not only do many people own a smartphone, but ownership is also prevalent across different social, cultural, age and gender variations, suggesting hard to reach individuals may be captured via their phones. Although recent surveys suggest there is a 'demographic digital divide' worldwide between young and old, those who are more or less educated and those with lower or higher incomes, rates of smartphone ownership are still relatively high in all but the least developed countries (Pew Research Centre, 2017; Poushter, 2016, p.11, p.16, p.20) In addition, it's important to consider that while ownership may remain low in some groups, the technological revolution is not slowing in scope or reach.

1.6.2 Smartphones support apps, which combined include many useful features and sensors for interventions

Mobile apps, or applications, are small software programmes with specialist functions that can be downloaded and installed specifically onto mobile devices such as smartphones. Native mobile applications are standalone programmes that are installed and run directly on the mobile device, occupying space in the devices memory. Web applications run through any web browser rather than from the device directly (e.g. Google Chrome). Some apps make use of data collected by smartphone sensors. Device sensors will vary, but could include accelerometers (often found in fitness trackers, they capture your movement and motion), gyroscopes (helps with orientation, for example if you tilt your screen), magnetometers (similar to a compass), Global Positioning Systems or GPS (receives information from satellites to pinpoint your location and is often used in map apps), not to mention more specialised sensors that might appear in some devices, for example many Samsung smartphones now incorporate a heart rate sensor (Forsblom, 2015). In relation to health, there are additional external sensors that can interface with mobile devices and apps, for example Fitbit's activity monitors (Fitbit, n.d.).

1.6.3 Smartphone apps allow for health behaviour monitoring, a popular technique for behaviour change

Tracking health behaviours is already a popular use for smartphone apps. Apps (and smartphones) have the ability to collect and collate a vast range of personalised data. Coined by staff working for Wired magazine in 2007, the phrase 'Quantified Self' or 'lifelogging' as it is also known, refers to *"self-knowledge through numbers"* (Quantified Self Institute, n.d.; Wolf, n.d.) or the idea that we can learn a lot by tracking and measuring things about ourselves. Evidence from a survey administered by The Pew Research Centre showed that 69% of American adult responders kept track of some sort of health indicator either for themselves or for a loved one (Fox and Duggan, 2013). There's also evidence that electronic tracking, for example via electronic diaries, is perceived as easier than paper-based tracking diaries (Marceau et al., 2007) and has featured in mobile health applications (if not specifically smartphone applications) as early as 2001 (Fogg, 2003, p.186). Lifelogging is a self-monitoring activity, for which there is an established evidence base for behaviour change (Dombrowski et al., 2012; Greaves et al., 2011; Harkin et al., 2016).

Shortly after this concept was defined, market leaders Google Play and the Apple App store opened (2008) and apps became widely available to the general public. Evidence suggests that many commercial health-related apps incorporate self-monitoring or tracking of behaviour (Bondaronek et al., 2018; Mendiola et al., 2015). Apps offer an easy, far reaching opportunity for members of the general public to learn about themselves by logging their behaviours in real-time and providing the potential for self-initiated behaviour change. Using an n-of-one design, apps could facilitate self-experimentation, allowing for data collection, analysis, and potential behaviour change or self-management (Choe et al., 2014; Taylor et al., 2018). As such, their potential for influence on health and wellbeing is substantial.

1.6.4 Smartphones and apps have been recognised for their potential for facilitating healthy behaviours and supporting health-care

Having a powerful computer at our fingertips, every day, has enabled changes to communication, education, finance, shopping and leisure among many other things. Naturally, it has also impacted healthcare and health management (Steinhubl et al., 2016). As B.J. Fogg writes, *"when you pack a mobile persuasive technology with you, you pack a source of influence"* (Fogg, 2003, p.186). Investigations into the potential of mobile phones for health promotion, began some time ago (Kaplan, 2006) and the term eHealth started to be used in academic circles in the early 2000s (Eysenbach, 2001). The World Health Organisation defines it as *"the use of information and communication technologies (ICT) for health"* (World Health Organisation, n.d.). With the availability of apps, and the apparent desire for individuals to self-monitor health indicators, it is perhaps unsurprising that a new and prevalent subsegment of eHealth emerged: mHealth (World Health Organisation, 2012).

Use of technology for healthcare is prevalent and growing. MHealth or mobile health, is *"the use of mobile and wireless technologies to support the achievement of health objectives"* (World Health Organisation, 2011, p. 1). This can include use of mobile and wireless features such as text messages, apps, or phone calls (Brinkel et al., 2014). MHealth approaches have or are being used in relation to a diverse range of health conditions and issues (e.g. Brinkel et al., 2014; Naslund et al., 2015; Parker et al., 2013; World Health Organization, 2011). Evidence is also available to suggest that mobile phones are highly acceptable for PA promotion (Monroe et al., 2015). The potential of health apps can be observed by the number and growth of the Health and Fitness app category in commercial app stores (Saadatfard and Årsand, 2016; Statista, 2018). While there is a digital divide within health apps (Bol et al., 2018; Carroll et al., 2017), the number of smartphone users with health apps is high, with one survey suggesting that a fifth of Northern American smartphone owners have such an app on their phone (Fox and Duggan, 2012). More importantly, some survey-based research has shown an association between health-app use and positive health behaviours and intentions (Sarcona et al., 2017).

1.6.5 Apps are a potential solution to some existing problems for physical activity promotion

Given their characteristics and popularity, smartphone apps have the potential to solve the problem of reach, adoption and maintenance based on their popularity and proliferation. They also have the potential to include effective and popular content such as self-monitoring.

Initial development costs can be high, but they are cheap to scale up (Turner-McGrievy et al., 2017). They will need maintaining, but this should not require large resources of time or people (definitive evidence could not be located but various web sources suggest maintenance costs approximately 15-20% of the development cost annually (Agicent, 2017)). Cost to the user is also cheap, with many apps being free or low-cost (e.g. 0.99p).

Finally, there is evidence for the benefits of self-management of health, and for health apps that facilitate self-management of health (Klein et al., 2014; Maes and Karoly, 2005). Given the shift towards self-management (or ‘Patient Activation’) in the health sector as advocated by the NHS Five Year

Forward View (National Health Service, 2014), apps are an ideal tool to support this. For PA, it seems plausible that apps could help facilitate wide-reaching small self-managed behaviour changes, which could result in the desired population shift advocated by Geoffrey Rose for public health intervention (Rose, 1985). This potential is being recognised more and more (Lathia et al., 2013; Pratt et al., 2012).

1.7 A possible solution: the motivating effects of acute affect

1.7.1 Distal or unobservable beneficial outcomes may not be sufficient motivators

As previously stated, PA is important for facilitating and maintaining good physical and mental health (Kim et al., 2012; Lee et al., 2012). Individually PA can help with changes such as weight loss, blood pressure, or on a whole-body scale, prevention of non-communicable diseases or conditions. Focusing on such health outcomes when encouraging behaviour change seems a *"logical, optimal motivator"* (Segar and Richardson, 2014, p. 840). Many PA interventions include informing participants of health benefits, or information on the consequences of their behaviour, typically via leaflets and often in conjunction with other intervention tools (Gardner and Rebar, 2019; Harland et al., 1999; King et al., 2008; Michie et al., 2009; Olander et al., 2013; Pears et al., 2016; Staten et al., 2004) .

A potential problem with using the health benefits of PA as motivators, is that many of them are difficult, or impossible for the individual to observe (for example, blood pressure, or prevention of a disease), and other benefits take a long time to emerge (for example, weight loss, or longer life). It's been proposed that these distal or unobservable outcomes may pose a problem for PA promotion (Biddle, 1992, p.179). As Bandura, (1991) posits *' Focusing on the more distal effects of courses of action...may provide little guidance for the future'*, meaning that individuals can't learn from such outcomes (p. 251). The decision to maintain a behaviour may depend on whether or not the individual is satisfied with what they have achieved by changing their behaviour (Rothman, 2000). If an individual can't 'see' the outcome or can't see it soon enough and subsequently perceive to have failed to achieve it, they may experience a loss of motivation (Guess, 2012). The perception of failure can also influence self-efficacy, an influential construct that appears in many behaviour change theories (Michie et al., 2014). A recent study showed that in-task feedback indicating success increased levels of self-efficacy, and self-efficacy decreased along with performance when feedback indicated failure (Achterkamp et al., 2015). Another showed that while longer term or unobservable goals such as healthy aging (e.g. pain free old-age), current health (e.g. cholesterol levels) and weight/appearance, were reported as reasons for exercising, those who exercised to improve *daily* quality of life – a shorter-term goal - reported performing more exercise (Segar et al., 2011). Researchers have started to recognise the need for short as well as long-term motivation for behaviour change, if only to optimise intervention design and prevent attrition (Bielik et al., 2012).

1.7.2 There are proximal benefits of physical activity, including positive affect

There are a number of immediate benefits to being physically active including improvements in working memory (Hogan and Carstensen, 2013), metabolism (Melby et al., 1993), and creativity (Oppezzo and Schwartz, 2014). However, individuals may be unaware of these benefits if they are in a real-world context. For example, the former and latter outcomes were demonstrated only in lab-based settings.

Despite a recent study being poorly interpreted as showing an association between high levels of physical activity and poor mental wellbeing (Chekroud et al., 2018; Stubbs et al., 2018), one established benefit of being physically active that could be identified by an individual in a real-world context is that it *"makes you feel good"* (Biddle and Mutrie, 2008; Ekkekakis et al., 2013; Fox, 2007, p. 413). The 'feel good' element can be operationalised in different ways. For example, it can reduce symptoms of clinical mental health disorders such as depression, relieve stress and provide feelings of achievement, enthusiasm, enjoyment or pleasantness. However, the immediate ‘feel good’ element of PA is thought to be short lived – i.e. it fades the further away the bout of exercise becomes in time (e.g. Bonham et al., 2018). The immediate feel good effect will be the focus of this thesis. Increases in positive affect can be recorded during or immediately after being active (Kwan and Bryan, 2010; Williams et al., 2008). Evidence suggests that a bout of as little as 10-15 minutes of PA can promote a change in affect/mood state (Darby et al., 2016; Ekkekakis et al., 2000; Focht, 2009; Hansen et al., 2001) and that the change in affect can last for at least three hours (Guérin et al., 2013).

1.7.3 Definitions of affect vary

The Circumplex Model (Russell et al., 1989; Yik et al., 2011) best describes this immediate ‘feel good’ element as part of core affect, and Russell, (2003, p. 148) elaborates, describing core affect as ‘*primitive, universal and simple’*. Core affect covers both positive and negative feelings, but not distinct moods (figures 2, 3). Researchers have supported the use of core affect to investigate the outcome of PA because it captures a range of positive and negative responses, typically using a single item measure, easy for repeat administration. In contrast, categorical measures of affect may identify specific mood states, (which are differentiated from affect as they are often longer-lasting with less obvious origins (Biddle and Mutrie, 2008, p.166)), better suited to research participants with clinical mental health disorders (see summary in Williams, 2008). Affect, rather than mood, is preferred for the current thesis as it is more relevant in the general population context (see figures 2 and 3 for further definition of affect relevant to this thesis).

1.7.4 Physical activity can influence affect through various mechanisms and variables

The complex way in which PA improves affect physiologically is beyond the scope of this thesis, but a number of suggestions have been discussed in the literature. Mechanisms include changes in neurotransmitters and endorphins

(Clow and Edmunds, 2014; Otto and Smits, 2011; Thoren et al., 1990; Tuson and Sinyor, 1993); exercise as a way to train the body to adapt to stressors (Otto and Smits, 2011) and physiological changes in brain-wave activity and the thermogenic hypothesis both demonstrate relaxation or reduced central nervous system arousal (see summary of evidence in Tuson & Sinyor, 1993, p. 97-104). For the purposes of this thesis, proof of change in affect induced by physiological changes is not required (although may be feasible e.g. Neale et al., 2017) and subjective changes in affect are acceptable.

To add to the complexity of this mechanism of effect are potential moderators and mediators of the physical activity-affect relationship such as duration of workout, level of exertion, gender (Rocheleau et al., 2004), self-paced/selected exercise intensity (Williams, 2008), BMI (Ekkekakis and Lind, 2006; Kanning et al., 2015), age (Schwaneberg et al., 2017), current activity levels/fitness (Bonham et al., 2018; Magnan et al., 2013; Parfitt and Hughes, 2009), type of exercise (Lane et al., 2005), being outside and being with people (Dunton et al., 2015; Thompson Coon et al., 2011), - including the idea of the bi-directionality of affect and exercise (Schöndube et al., 2016), among many others (e.g Annesi and Westcott, 2007). In addition, there is also the hypothesis that the affective benefits of PA are due to a placebo effect, whereby individuals believe exercise improves affect and so it does (Desharnais et al., 1993).

As Emerson and Williams (2015) note, there are multiple potential moderators and mediators and 'physical activity feels good' is an insufficient descriptor for this complex relationship. One model that may help explain some of the aforementioned moderators and mediators is the dual-mode model. It suggests that cognitive factors and physiological cues together specifically influence exercise-induced affect (Ekkekakis, 2003). The influence of these two factors is thought to change based on the intensity of the exercise being undertaken, with lower intensities resulting in stronger cognitive influences, and higher intensities resulting in physiological cue influences.

Figure 2 Defining affect 1

In the context of this thesis, the term affect refers to short-lived feelings or *'the [current] conscious experience'*, both positive and negative (Russell, 2003, p. 148). The Circumplex model suggests that there are two dimensions to affect: pleasure-displeasure and arousal-sleepiness and your state can be anywhere along the dimensions but both will feature. For example you can be excited (high pleasure and arousal), feel stressed (low pleasure and high arousal), non-clinically depressed (low pleasure and sleepy) or relaxed (high pleasure and sleepy) (Russell et al., 1989). This has subsequently been expanded to a 12 point model with additional items such as feeling frenzied, distressed, gloomy, sluggish, tranquil, peaceful, enthusiastic and energetic (Yik et al., 2011), see below.

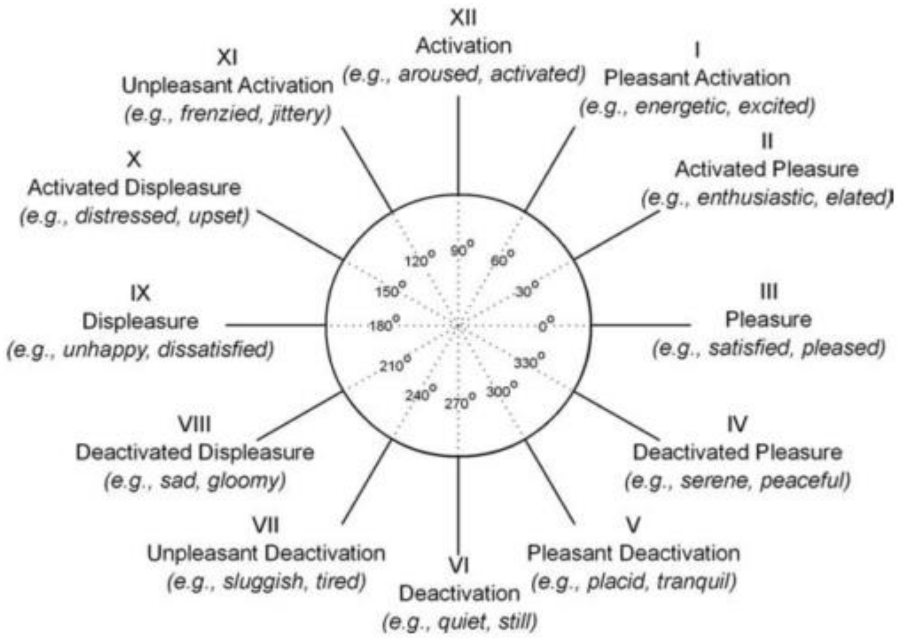


Figure of 12-PAC from Yik et al., 2011, p. 706

Figure 3 Defining affect 2

Non-clinical affect will be the focus of this thesis. This is because clinical affective disorders require specialist treatment which is outside the scope of this work. Although there is evidence to demonstrate the impact of physical activity on clinical affective disorders, equivalent to the impact of medication, the focus is not on improving mental health, but improving physical activity levels. Therefore, when referring to affect it will be non-clinical.

Although affect is thought to be a simple, state-based component of emotion and mood, (Yik et al., 2011, Russell 2003), with complex emotion taking longer to form (Baumeister et al 2007) there has not always been a distinction made, or clarity provided between these three terms (e.g. see Russell 2003), and the literature can be confusing. Therefore, where necessary varied terminology may be allowed (see methods and results chapters).

Hedonic wellbeing (discussed later) is different to eudaimonic wellbeing - *"the state of personal well-being in a holistic sense",* according to Aristotle (Ekkekakis and Dafermos, 2012, p. 303) – in that it looks at increased pleasure and decreased pain to lead to happiness and measures subjective well-being, whereas eudaimonia is related more to the idea of self-realisation. Again, where necessary, both will be allowed due to the often overlapping use of relevant terminology and measures such as 'happiness' (often related to hedonia) and 'vitality' (often related to eudaimonia) as well as the idea that the two are linked (Ryan and Deci, 2001).

The 'feel good' element will be primarily described as 'positive affect' for the purposes of this research, and will focus on 'simple' feelings (affects) associated with bouts of PA such as feeling upbeat, enjoyment, feeling better and so on (e.g. Lox et al. 2000; Hardy & Rejeski 1989; Laverie 1998).

1.7.5 Recognition of affect as a factor in behaviour change is rising again

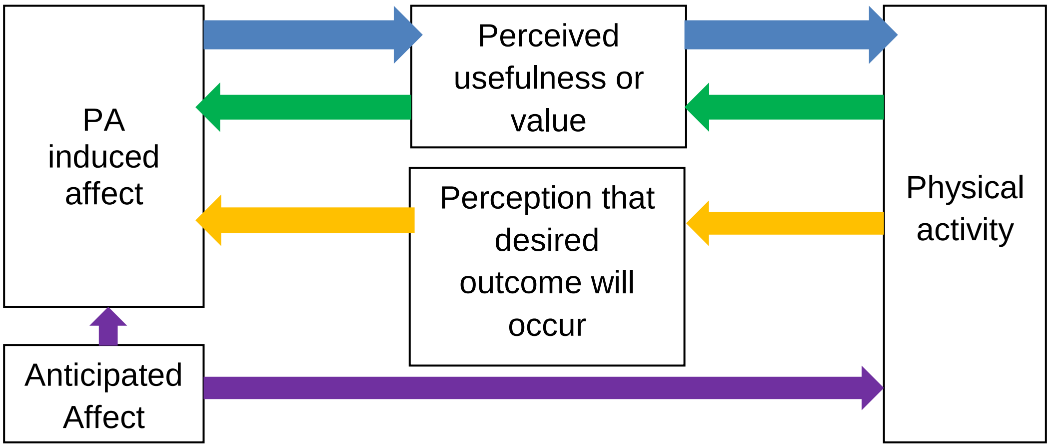
While the relationship between PA and affect is likely very complex, so too is the more pertinent relationship between affect and subsequent PA levels, this too having potential mediators and moderators such as psychological need satisfaction and self-efficacy among others (McAuley et al., 2003; Schneider and Kwan, 2013). Earlier theories and models of behaviour change considered humans as rational beings, who made informed decisions. The poor uptake of PA and other health behaviours in the face of vast and various methods of promotion suggest that such thinking is outdated (Ekkekakis, 2017). With previously limited research investigating the motivational features of affect (Biddle, 1992, 1995; Gauvin and Brawley, 1993), it is subsequently experiencing a resurgence, especially for PA research (Ekkekakis et al., 2013; Ekkekakis and Dafermos, 2012; Williams, 2008). Recent discussions in the literature call for the inclusion of affect as a potential influencer for behaviours and a move away from focusing on purely cognitive mechanisms (Ekkekakis, 2017; Jekauc and Brand, 2017; Lewis et al., 2017).

1.7.6 There are many theoretical explanations of the relationship between affect and subsequent physical activity

There are now theories that focus entirely on the role affect plays in behaviour. Williams (2008) provides an excellent summary of one of these collections of theories: hedonic theory. Hedonic theory, or the hedonic principle, has a long history with discussions of the influence of pleasure or affect appearing in scientific texts from as early as 1789. Despite the terminology used, it is not limited to a single theory, rather it is the hypothesis that *"behaviour is a function of its affective consequences or anticipation of its affective consequences"* (Williams, 2008, p. 4). Many theories draw on this hypothesis and therefore, the exact mechanism of effect becomes complicated. In all likelihood it is a combination and interaction of the proposed mechanisms described in the various theories.

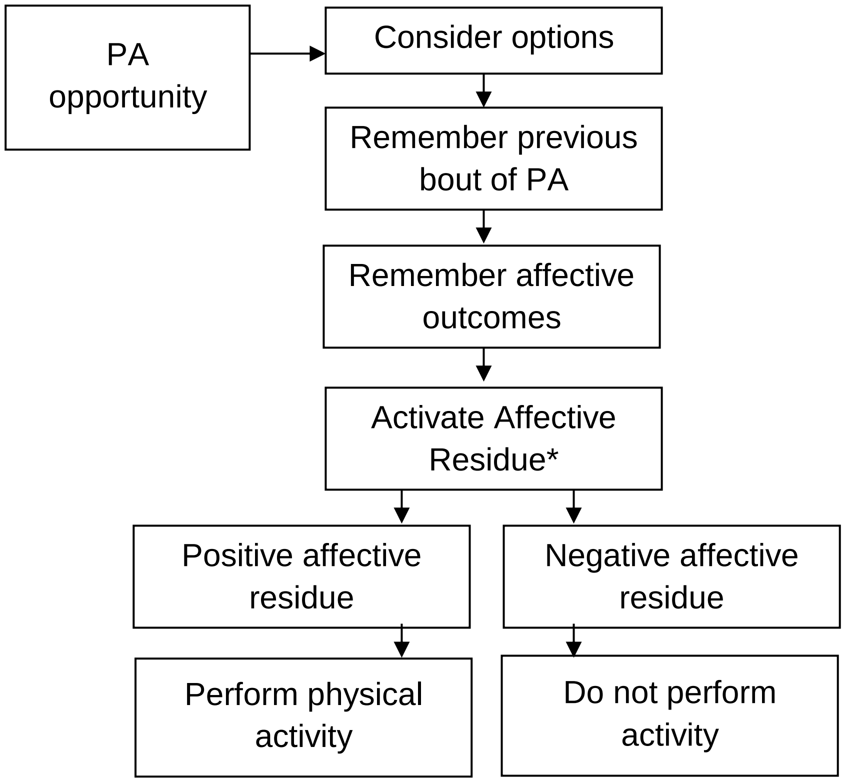
Kahneman, for example, adheres to the theory that the affective response to a behaviour defines its perceived usefulness, and thus whether it will be performed again, (Williams, 2008). Other theories look at anticipated affect (AA) a type of affective attitude, defined as *'the affect that is expected to be experienced'* (Sala et al., 2016). The Response Expectancy Theory and the Expected Pleasure Theory both suggest that AA determines enactment of a behaviour and also propose a link between AA and actual affective response. The latter theory suggests that the actual affective response to a behavioural outcome is dependent on the individual's perception that the outcome will occur, as well as the perceived value of the outcome (see figure 4 for an example of how these factors may interact to influence PA as described). Here we can begin to understand how a difficult to observe physiological outcome such as weight loss or reduced blood pressure, could result in low affect if the value is perceived as high (e.g. weight loss), but the perception of the outcome occurring is low (e.g. weight loss doesn't happen quickly, so it could be perceived as unlikely to occur). Subsequently AA could be low, resulting in reduced engagement with the programme of exercise.

Figure 4 Potential mechanisms of effect between affect and physical activity



Baumeister et al., (2007) uses the hedonic principle to suggest that emotion can work both directly and indirectly to influence behaviour. The direct approach is supported by a range of theories, including in particular, evolutionary and social theories, and suggests that emotion 'drives' active behaviour for the betterment of the individual. An example being causing arousal to better assist in fight or flight behaviours. In contrast, the indirect pathway proposes that previous and current affect influence behaviour as a feedback mechanism (see figure 5 for a simplified illustration of the indirect pathway). The theory appears related to principles of associative learning (e.g. operant conditioning) and the law of effect, with a behaviour (physical activity) and a consequence (affect) becoming associated, prompting that behaviour to be performed (or not) in future (Skinner, 1938; Thorndike, 1898).

Figure 5 Illustration of indirect pathway between affect and physical activity



*Stored left-over affect from previously performing the behaviour

Williams (2008) proposed his own model of how exercise-induced affect could influence future behaviour, by combining both hedonic principles and the dual-mode model to incorporate affective, cognitive and physiological influences (see figure 6).

Figure 6 Williams 2008 model of self-paced exercise, affective response, and exercise adherence

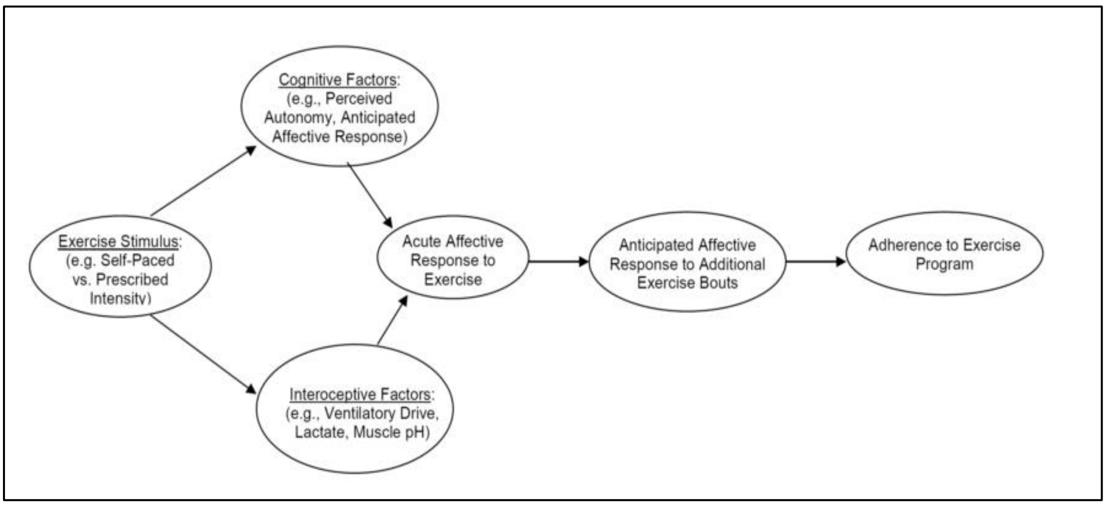


Figure reproduced from (Williams, 2008)

Theories began changing or being developed to explore affect in the 1980's (see summary in introduction of Lox et al., 2000). There are now multiple theories of behaviour change that include some facet of affect either as a construct in its own right, or as part of a construct. Michie et al., (2014) describe 83 behaviour change theories. A key word search indicated that at least 30 of them mention the influence of affect, mood or emotion. Social Cognitive Theory (Bandura, 1986, 1991) now includes a mood-related sub-function, suggesting that mood is involved in self-monitoring of behaviours, and can impact perceptions of efficacy and recall of the behaviour, as well as facilitate identification of patterns. The Self-Determination Theory suggests affective responses influence intrinsic motivation (motivation derived from inherent interest or enjoyment) (Ekkekakis et al., 2005). The Theory of Reasoned Action has incorporated mood and emotion as a potential influence on beliefs (Fishbein and Ajzen, 2010). Its successor, the Theory of Planned Behaviour, also features Affective Attitudes (AAt) and evidence suggests they are predictors of exercise-identity profiles, with AAt able to predict those who intended to exercise and those with strong exercise identities (De Bruijn et al., 2012). In addition to these theories, there has also been an acceptance that mood states are influential in self-regulation (e.g. Hall and Fong, 2007) and relapse prevention or maintenance (e.g. Hendershot et al., 2011). Many other theories propose other potential pathways, draw on the hedonic principle, or are under development (e.g. Van Cappellen et al., 2017). To report on all of them is beyond the scope of this thesis.

The current overview of the evidence and some of these theories, should serve to illustrate that while the method of influence of affect is varied and not necessarily agreed upon, a consensus is starting to form among the research community of a circular relationship between affect and behaviours - with behaviours such as PA in particular influencing affect, and affect (in general or specific to the behaviour) influencing behaviour (e.g. Schwerdtfeger et al., 2010).

1.8 What evidence is there that affect might promote physical activity?

1.8.1 There's a variety of evidence that shows affect could influence physical activity levels or intentions to be active

Irrespective of how PA improves affect and how affect influences behaviour, there is evidence to suggest that positive affect in general can predict or is associated with future participation in PA or intentions to be active. For example, general affect is associated with subsequent free living PA (Garcia and Archer, 2014; Liao et al., 2015; Schöndube et al., 2016) and affective evaluations of exercise are associated with subsequent activity (Brand and Antoniewicz, 2011). Affective judgements (*'reflections or expectations about the overall pleasure/displeasure, enjoyment and feeling states expected from enacting [a behaviour]'* (Rhodes et al., 2019, p. 19)), anticipatory affective gains (knowledge of the affective outcomes from being active), and interventions that seek to manipulate them are associated with PA behaviour change, although only a small-to-medium effect size was found for the former (Rhodes et al., 2019; Rhodes and Quinlan, 2015; Sirriyeh et al., 2010).

When looking at PA-induced affect specifically, there is evidence of an association both in inactive, or not completely inactive, adults (Kwan and Bryan, 2010; Liao, 2015; Williams et al., 2008, 2012). One study also reported that an increase in positive affect induced by a walk as brief as 10 minutes was strongly and positively correlated with future intentions to be active in sedentary, obese women (Focht, 2013). Although more important to active individuals, qualitative and survey data suggest that positive affect is an acknowledged motivator for participation in PA (Aaltonen, Kujala, et al., 2014; Aaltonen, Rottensteiner, et al., 2014; Hardy and Grogan, 2009; Laverie, 1998), while inactive students have requested PA promotion apps with the ability to monitor feelings (Middelweerd et al., 2015). A recent review of the literature reported that positive affect experienced during exercise was associated with subsequent PA, but failed to find an effect for post-exercise affect (Rhodes and Kates, 2015).

1.8.2 There are limitations and gaps in the literature

Evidence comes predominantly from lab-based studies of structured exercise-induced positive affect (e.g. running on a treadmill). These are useful for proof of concept work demonstrating the effect in controlled conditions, but may not reflect real-world experiences. In addition, such structured activities may be less favourable to inactive individuals for a range of reasons such as intensity and type, as well as social aspects such as being watched by other exercisers or feeling negative about their body. It's less clear whether there is a body of literature demonstrating the impact of positive affect induced by free-living PA (e.g. walking as active travel or for leisure) on free-living PA levels and intentions.

In addition and of importance, is that many of the studies were not looking to promote PA. They were looking to identify and clarify the hypothesised relationship between physical activity-induced affect and subsequent PA. Therefore it could be argued that the observed association between PA-induced immediate affect and PA in a free-living environment is an unexplored avenue for experimental research into PA promotion.

1.8.3 Using affect to promote physical activity may require a range of techniques

The acute affective benefits of PA are not necessarily as well known as the physical benefits. Although some evidence suggests that the public are aware of the affective benefits associated with being active, (Murray, 2006; Reavley and Jorm, 2012), this awareness does not appear to be sufficient to motivate behaviour change. In addition, individuals from more deprived areas may be less convinced of the benefits of PA (Murray, 2006) and while those of lower socioeconomic status may be less aware of recommended levels of activity to begin with, let alone benefits (Knox, Esliger, et al., 2013), the use of such thresholds in promotion may not be beneficial for understanding its health benefits anyway (Knox, Webb, et al., 2013). Evidence suggests that positive perceptions of the health benefits of PA can help moderate perception of barriers to being active and facilitate exercise (Mcguire et al., 2016). Therefore it seems that in the first instance awareness of the links between activity and affect need to be raised (as advocated in the literature

(Carels et al., 2007)), and it needs to be done in a way that is sufficient to promote behaviour – not just measure behaviour and affect.

Evidence suggests that self-monitoring of behaviour/goal attainment is a good technique for changing behaviours, with a large evidence base (Dombrowski et al., 2012; Greaves et al., 2011; Harkin et al., 2016). Bandura, (1991) describes self-monitoring and its characteristics such as "*fidelity, consistency and temporal proximity"* as an important part of self-regulation for behaviour change (p.250). There is also evidence that self-monitoring is an effective and acceptable way to raise awareness. One study found that it made people more aware of their health condition, as well as able to identify a cause and effect relationship between their behaviours and health and participants were primarily satisfied with the intervention because the self-monitoring generated self-awareness (Tomita et al., 2008). Another has shown that self-monitoring through an app also raised awareness of PA levels (Buman et al., 2016). Self-monitoring also allows for an element of personalisation, or tailoring – the data is yours and describes you and your experiences alone. Tailoring is also thought to be associated with behaviour change (e.g. Bull et al., 1999; Ghanvatkar et al., 2019; Short et al., 2011).

In addition to self-monitoring, feedback on behaviours and their outcomes can be effective for raising self-awareness and inducing behaviour change or goal attainment (see summary in Colineau and Paris, 2011; see app in Consolvo et al., 2008; see app in Maitland et al., 2007) with some evidence to suggest progress-monitoring combined with immediate feedback is even more effective (Harkin et al., 2016). Users have also reported real-time feedback (as well as monitoring and other features) within apps as an important feature for their engagement with PA (Bort-Roig et al., 2014). These techniques can facilitate uptake (Samdal et al., 2017) as well as maintenance (Dombrowski et al., 2012; Samdal et al., 2017), as evidence suggests that past behaviour can predict future behaviour (e.g. Ouellette and Wood, 1998).

Informed by learning from behaviour change experts and studies mentioned above, if the association between exercise and exercise-induced acute affect could be highlighted to participants, using their own personal experiences, it may serve to promote future activity levels. Like early self-quantification research suggests (see summary in introduction of Maltseva and Lutz, 2018), one promising mode of delivery that is already popular for self-monitoring, and could enable measurement of PA and affect as well as personalised feedback, potentially highlighting the association, is the aforementioned smartphone app. This results in the first two questions for this thesis:

RQ1: Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?

RQ2: What are the characteristics and content of physical activity apps that include feedback on immediate affect, including both apps developed for/by researchers and publicly available apps (commerical apps) in the app stores?

1.9 What evidence is there for physical activity apps that provide feedback on affect?

1.9.1 There are many physical activity promotion apps, but it's less clear how many apps there are that provide feedback on affect

Apps for PA promotion are multitudinous. A review of the iTunes App store (which is now the Apple App Store) from 2012, reported that PA, personal health and wellness and healthy eating apps were more common than apps for substance abuse, emotional health, safety and sexual/reproductive health (West et al., 2012). Orcha, a website that reviews health apps, reports that as of 2018 there are approximately 327,000 apps in the health and fitness category across app stores (www.orcha.co.uk), therefore the number of PA promotion apps either focusing only on this behaviour, or included as part of a lifestyle app, is likely to be very high.

Feedback on happiness experienced when eating has been used in a recent app (Renner et al., 2018; Wahl et al., 2017), but evidence of PA apps with this function is less clear and no review is available. A scope of the literature suggests there may be some that exist. One mHealth intervention used motivational messages citing the affective benefits of exercise, however these weren't based on users' own affect, but a generic understanding that affect could be improved by exercise (Martin et al., 2015). Earlier work describes the Personalised Information Platform for Health and Life Services (PIPS) study, which used mobile phones to deliver a PA intervention to diabetic patients (Erriquez and Grasso, 2008; Morandi and Serafin, 2007). The intervention collected information on the users' emotional status (among other things) in order to tailor motivational messages they subsequently received. The messages could then include personalised references to their activity behaviour in relation to their affect. Such personalisation aimed to raise awareness and allow for behaviour change. The Wish Outcome Obstacle Plan app (The WOOP App, n.d.) is currently available in stores, and was developed and tested by researchers using established behaviour change strategies as well as reporting an extensive evidence base for behaviour change (e.g. Stadler et al., 2009). It asks users among other things, to define the best outcome for themselves, including how *'fulfilling your wish would make you feel'*. There has been a call to use an app to focus on the immediate personal affective benefits of exercise, and use this feedback to encourage future exercise (Stevens and Bryan, 2012) and while of interest, the aforementioned apps do not fill that gap. One used generic affect rather than the user's affective response to exercise, and the other failed to target individuals without a clinical condition (making large-scale population change difficult as this is a specialist group). This reinforces the need for the first research question:

RQ1: Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?

1.9.2 Evidence for the effectiveness of physical activity apps is modest and may be hampered by usage

While evidence for PA apps providing feedback on affect is limited, there has been a lot of research into PA apps in general. A number of reviews exist that explore the effectiveness of apps for behaviour change and suggest that while promising, evidence is still modest at best.

Some reviews report findings in support of the use of mobile devices in general for PA promotion, (Fanning et al., 2012; Muntaner et al., 2015). Smartphone apps have also been assessed alongside text messaging interventions, but they were still in the minority and no effect for PA was observed (Stephens and Allen, 2013). Other reviews looked at apps independently, with some reporting only modest or limited effects of apps for PA promotion (Bort-Roig et al., 2014; Coughlin et al., 2016; Martínez-García et al., 2017) and another reporting evidence for a non-significant improvement in PA (Mateo et al., 2015). Finally, one review suggested that there is evidence for their effectiveness across a range of behaviours (Zhao et al., 2016). This included one app that focused on increasing PA alone and two that included PA promotion as part of a general lifestyle intervention.

Effectiveness may be impacted by usage. If public apps are being developed for profit, it’s likely that usage is a key priority for developers. As common-sense would predict, there is evidence that usage of a digital intervention (in this case a website) is linked with intervention effectiveness (Alexander et al., 2010) and that usage can be influenced by many intervention components (see text messaging intervention by Redfern et al., 2016). Usage, or engagement, with digital interventions is thought to be a complex process, with multiple facets including temporal, behavioural, experiential and state-like characteristics (Perski et al., 2017), making it easy to see how effectiveness could be affected by how it is promoted or manipulated within an app. In addition, measurement of user engagement or usage has previously been limited, calling for new ways to measure and report these outcomes (Miller et al., 2018).

1.9.3 The quality of physical activity apps is questionable and assessments of quality appear limited to certain characteristics

The quality of public apps for behaviour change has started to be evaluated by researchers and the consensus is that there is still much to be desired. For example, reviews of weight management and PA apps concluded that there is minimal use of evidence-based strategies or public health exercise targets, minimal theoretical content, apps are of average quality and information quality was typically poor (Bardus et al., 2016; Breton et al., 2011; Cowan et al., 2013; Knight et al., 2015; Pagoto et al., 2013). In contrast, there's evidence that PA apps tended to use the same behaviour change techniques (BCTs) used most frequently in other PA interventions, suggesting a degree of evidence-based content (Middelweerd et al., 2014). However, another review found that PA and dietary apps that included BCTs associated with effectiveness were more likely to feature in paid apps (Direito et al., 2014), suggesting along with Cowan et al., (2013) that better quality apps may only be available to those willing and able to afford them. Aside from content are also the questions of privacy, development quality and usability. Although fewer reviews examine these and findings are not overly positive, they are starting to gain recognition as factors that could influence intervention success (Bielik et al., 2012; Bondaronek et al., 2018). This leads to the third research question:

RQ3: What is the quality of physical activity apps that provide feedback on immediate affect?

There are now recommendations to focus more on such elements of quality in general (Mateo et al., 2015; Monroe et al., 2015; Roberts et al., 2017; Rose et al., 2017). However, as Freeman et al., (2017) suggest in their pleasingly titled paper 'Why the public health sector couldn't create Pokémon Go', (an app-based game where users were inadvertently active due to the need to travel to catch mythical animals), researchers and public health officials are likely to come at app design from a very different perspective to app developers, take longer to develop an app, use techniques that may have reduced in popularity by the time efficacy is established and advertise the app poorly. Even Public Health England now recognise and encourage multi-disciplinary approaches when applying behavioural and social sciences to improving health (Public Health England, 2018). Therefore, it is important to learn from both the existing literature (involving behaviour change experts) as well as existing commercial (public) apps (involving expert developers) to determine the characteristics and quality of the apps of interest in order to make informed and comprehensive recommendations. To that end, both a systematic review of the existing literature and an evaluation of existing publicly available apps available in app stores will be conducted in order to address the compiled research aims, questions and objectives.

1.10 Aims, research questions and objectives

Aim: To assess the characteristics and quality of apps that use feedback on affect to promote physical activity

Research questions:

1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps that include feedback on immediate affect, including both apps developed for/by researchers and publicly available apps (commercial apps) in the app stores?
3. What is the quality of these physical activity apps that provide feedback on immediate affect?

Objectives:

1. A systematic review of literature of physical activity smartphone apps that include feedback on affect will be conducted to determine evidence from a research context (which should prioritise evidence-based content).
2. An evaluation of publicly available smartphone apps will be conducted to determine evidence from a real-world context (which typically

prioritises user engagement).

3. A set of recommendations will be made for the future development of an app-based intervention to promote physical activity, using feedback on affect.

2. Methodological approach

2.1 Introduction to chapter

The following chapter will outline the paradigmatic standpoint – pragmatism - for the thesis. It includes a consideration of ontological and epistemological beliefs. Strengths and weaknesses of pragmatism are discussed and an argument made for its selection. The use of both quantitative and qualitative methods, within a mixed methods research approach is discussed. An argument is made for combining approaches. Methodological choices for the studies included in this thesis are addressed, as well as data collection methods and the reliability and validity of those methods. Further discussion of their development or adaptation (where appropriate) is in subsequent Methods chapters. More detailed discussion of the bias, reliability and validity of study procedures and data collection methods is included in Methods chapters.

2.2 Paradigm, Ontology and Epistemology

2.2.1 Terminology

Paradigm (or theoretical perspective), ontology and epistemology are often used and/or defined, interchangeably within the research literature. Therefore, for the purposes of this thesis, the model depicted in figure 7, modified and informed from Guba and Lincoln, (1994) and Teddlie and Tashakkori, (2009), shall be adhered to. Here it is proposed that the paradigm and its characteristics should be reported at the outset and be defined by its ontological and epistemological stance.

Figure 7 Flow diagram of philosophical concepts for research

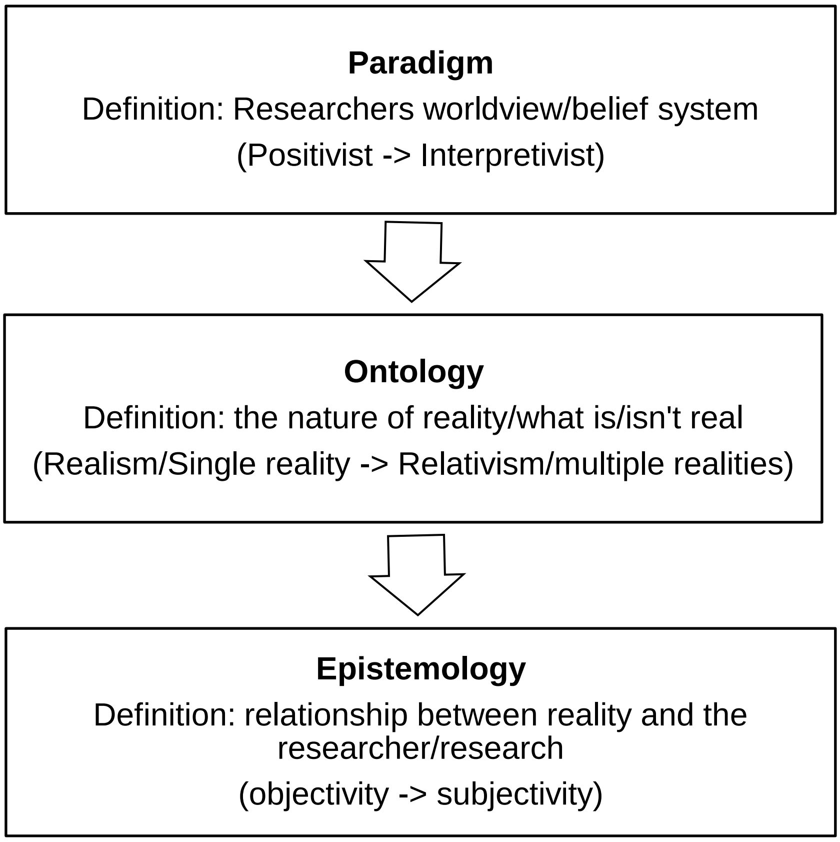


Figure informed by table 5.2 in Teddlie & Tashakkori (2009) and Guba & Lincoln (1994)

Crotty, (1998a) states that four things need to be established when considering a programme of research: 1) methods, 2) methodology 3) theoretical perspective (paradigm) and 4) epistemology. The concept of ontology should also be added: *'the nature of reality'*, or what is or isn't real (Lincoln & Guba,1985, p. 37). Using a modification of Crotty's approach (1998a), the paradigm will be established first, as this reveals the researchers 'worldview' which *'guide[s] disciplined inquiry'* (Guba, 1990, p. 18), followed by ontological and epistemological characteristics and finally, methodology. Methods will be reported in Chapters 3 and 4.

Broadly, paradigmatic assumptions relevant to research sit on a continuum. At one end are those who believe that reality is made up of measurable objects that can be tested and exist even when humans are not interacting with them. For example, Positivists would consider that a chair is still a chair, whether or not they are sitting on it. At the other end of the spectrum are those who believe reality is made up of subjective experiences – how humans react to, and perceive their conscious surroundings. For example a chair is only a chair when someone sits on it, but when someone stands on it to reach something, it's a ladder (O'Gorman and MacIntosh, 2012). This is an Interpretivist stance. Other paradigms exist between these two poles (Mertens et al., 2010; Teddlie and Tashakkori, 2009).

The two extreme paradigms (positivist and interpretivist) can be parsed into a variety of approaches based on nuanced assumptions about reality (Guba, 1990). Philosophers have identified four ontological perspectives that fit under and along the paradigmatic continuum: objective reality, perceived reality, constructed reality and created reality (Lincoln & Guba, 1985, p. 82-87). As Lincoln and Guba (1985) state, in objective reality, 'naïve realists' consider the whole the sum of the parts and that once enough research is done, an answer will be converged upon (p.83). The perceptual or critical realists suggest that while reality is 'out there' it can never be known in full – the whole is more than the sum of the parts (p.83). Constructionists, or relativists, are unconvinced that a single reality exists, instead favouring the idea of multiple realities. If reality is constructed within the mind there is no single whole, the parts allude to different wholes (p. 83-85). Finally, creative realists don't believe a particular reality exists. Instead there is the potential for different realities where, depending on what happens and who participates, a particular reality is brought into being – there is no definitive whole and the parts dictate which whole comes into being (p. 85-87).

Understandings of epistemology – '*the relationship of knower to known'* (Lincoln and Guba, 1985, p. 37), or the '*grounds of knowledge/relationship between reality and research'* (Carson, Gilmore, Perry, & Gronhaug, 2001, p.6) also exist on a spectrum. At one end sits the idea that one can obtain definitive knowledge in an objective manner, i.e. the data was and generally could be collected independent of the particular researcher. Lincoln and Guba (1985) describe the researcher and the thing being measured as independent, referred to as 'subject-object dualism' (p.93). In contrast, is the epistemological standpoint that the researcher and the knowledge are interlinked and therefore can influence each other, leading to the knowledge being subjectively obtained. This could be referred to as subject-object unity, or the idea that reality is in the mind.

2.2.2 Thesis Paradigm/Theoretical perspective

One paradigm not mentioned above was pragmatism. This is because, unlike most of the other paradigms, it tends to reject the need to choose a single end of the 'reality spectrum'. According to Guba & Lincoln (1994), choosing your research paradigm is the most important part of determining your subsequent methodology and therefore methods. However, in contrast to this focus on the importance of paradigm, pragmatists forgo the use of a single worldview and set of assumptions about reality and instead prefer to emphasise a '*practical approach…to research problems'* (Denscombe, 2010, p.138). Instead of looking at what is or isn't real (ontology) or how we can know things (epistemology), pragmatists focus on how useful the knowledge is when applied to a practical problem (Denscombe, 2010).

A pragmatic approach therefore is to let the research questions inform the methods, to ensure that the knowledge being derived is of most use for answering the questions. As this thesis will examine characteristics and quality of PA apps, and hypothesises that quality can, and should be defined in various ways relating to both ends of the reality spectrum, a pragmatic approach is called for. While a positivist stance will be relevant to the identification of the few established quality indicators in apps and behaviour change interventions, interpretivist stances are relevant for determining the perceptions of potential users. Therefore the 'third way', pragmatism, that recognises the need to apply a paradigm on a need by need basis, is most appropriate (Armitage, 2007).

2.2.3 Thesis Ontology

The pragmatic paradigm allows for two preferred ontologies to be applied, based on the target areas of interest for this thesis. Existence, characteristics and quality of apps will be examined using both critical realist and relativist interpretations.

The extreme stance of positivism has many acknowledged weaknesses and has been superseded by the more flexible and preferred, post-positivism (e.g. Clark, 1998; Gray, 2018). The post-positivist paradigm tends to align with the critical realism ontology. Critical realism fits with the reality of the quality of apps, as described in the current literature, namely that there are some demonstrated associations between certain app variables (e.g. presence of evidence-based content) and high quality, as well as certain quality variables and effectiveness (e.g. behaviour change techniques). However, the literature remains imperfect, incomplete, and is unlikely to be able to explain all possible variance. Therefore, the relativist ontology will also be used to attempt to identify those app features that are perceived and constructed as facilitating or inhibiting use and behaviour change for and by individuals. This ontology has been chosen because it fits with the idea that app users will create their own reality when choosing and engaging with an app, based on their own intentions and motives for use.

2.2.4 Thesis Epistemology

John Dewey, an early advocate of pragmatism, proposed that it is important to recognise that dual-direction interactions were happening in the world. The individual acts upon the world, and the changes this makes in the world then act upon the individual. It is the relationship between actions and consequences that is key for knowledge accrual, rather than objective or subjective reality (Biesta, 2010). Therefore, like the pragmatist ontology, the pragmatist epistemology also allows for a bridging of the gap between objectivity and subjectivity in relation to the researcher and their relationship with knowledge. In order to understand the relationship between actions (inferred cause) and consequences (inferred effect), pragmatists acknowledge that at times, interaction between researcher and participants may be required (subjective interpretations of both causes and effects) while at other times, interaction may not be required (objective interpretations, or identification of causes and effects) (Teddlie & Tashakkori, 2009, p.90). In this thesis both objective and subjective interpretations will be required – both to determine presence of evidence-based content and determine subjective indicators of quality- possible causes and effects related to uptake and maintenance.

2.2.5 Strengths and weakness of pragmatism

While it can be argued that the main strength of pragmatism is its flexibility to adapt to the research question under consideration, this means that unlike other paradigms, pragmatists will be at risk of any and all biases associated with their approach to answering the question. Weaknesses of both critical realism and relativism and objective and subjective approaches will be relevant (see table 1 for examples). However, as the idea behind pragmatism is that the variety of options available means that selections can complement each other, this should ideally lead to a strengthening of outcomes rather than a weakening. For example, findings should or could triangulate (where appropriate), or in the case of this thesis, answer different parts of the same question, providing a more complex picture and strengthening any argument.

Table 1 Examples of strengths and weaknesses of pragmatism and its related ontological and epistemological approaches

Approach	Strengths/Benefits	Weaknesses/Disadvantages
Pragmatism	Flexible investigative techniques. More likely to promote collaboration between researchers. Can use qualitative research to inform quantitative research and vice versa. Can use qualitative research to validate or explore further quantitative research and vice versa. Including quantitative data allows for the fact that qualitative data cannot be generalised. Including qualitative data can explain relationships discovered in quantitative data. Able to combine empirical and descriptive precision. Ability to combine macro and micro levels of a research issue.	Perceived as less concerned with rigorous scientific or intellectual discipline. Combining qualitative and quantitative methods can be seen as inappropriate – they are incompatible. Therefore findings are merely assigned truth but may not represent truth. Doesn’t necessarily have a theory of truth – unlike other approaches.
Critical Realism	It recognises that phenomena may be defined by a set of practices as well as ideas about those practices, which may or may not be correct – therefore allowing for an analysis of the social world as well as the natural world. Recognises the fallibility of scientific enquiry and	The social world being as it is means that there are limits to what scientific processes can achieve. Belief that if an ‘intelligible’ occurrence happens it infers what the world is like – critics suggest that this merely implies what scientists believe the world is like, not

	recognises the social nature of knowledge development (unlike Positivism). Recognises that constructivist paradigms may focus too heavily on human perspectives and cannot always account for knowledge development. As such suggests three realms of reality that account for different objective and subjective perspectives – attempting to reconcile agency and structural factors. Well suited to responding to research questions that relate to understanding complexity – advocates understanding phenomena in the real-world, not necessarily under controlled circumstances. Seeks to explain and understand phenomena rather than just observe it, and so does not inherently prescribe use of specific methods – flexible.	necessarily what it is actually like. CR may not be the only theory about scientific practices that suggest scientific activities are ‘intelligible’. Conclusions from scientific inquiry still depend on the scientific knowledge of the day and therefore may not reflect the true nature of the world (although CR is thought to acknowledge this limitation and the fallibility of its position). Its critique of traditional methods of empirical testing means that possible explanation of a phenomena may be difficult to test for. The idea that truth is supposed to correspond to objects in the real world, doesn’t leave much flexibility for competing explanations of a phenomena and means that it’s difficult to decide between them if there are ‘multiple truths’ . In fact there are multiple types of truths proposed in this theory. The perception that it’s neither ‘critical’ enough as its limited to a critique of methodology, rather than theory, nor ‘realist’ enough as it does not understand core phenomena of various domains.
Relativism	Aims to represent a range of realities from a range of people.	The idea that Relativism is self-contradictory, as it makes a nonrelative statement that all things are relative.

	Adept at handling social and behavioural realities. Focuses on understanding, not explaining phenomenon, thought to be more appropriate for human sciences. Quality of data is determined by credibility - a consensus between informed/qualified people. This may be easier to reach in instances where empirical experimentation is inappropriate or unethical. Provides transferability as a sufficient equivalent of generalisability in the positivist paradigm. Research takes place in natural environments, where the phenomenon occurs, meaning results are perceived as authentic.	Fallibility – If there are no universal truths, then relativists can never know for sure if what is observed represents the truth. This means it's difficult to make judgements among different claims to knowledge, especially if the empirical method of looking at the quality of the methods employed is not deemed acceptable. Tends towards credibility of outcomes, rather than validity of outcomes.
Objective/ Objectivism	It's a simple perspective – knowledge approximates the reality of phenomena, or to put it colloquially – what you see is what you get. Provides information about the phenomenon itself.	Provides information about the phenomenon itself but not necessarily its meaning. Assumes researchers can detach from their values and emotions and not introduce bias. However, even 'objective' measures of phenomena have been developed by humans, making decisions that are open to bias.
Subjective/ Subjectivism	Emphasises the role human perceptions and perspectives play on behaviours. Doesn't reduce people and their response merely to social structures.	Suggests the world is truly unknowable, and therefore may disregard whether the subjective impression corresponds to any other notion of reality – i.e. other data may be ignored.

		Methodology and validity are irrelevant. Social and natural influences can be overlooked. Provides information from a variety of viewpoints on a phenomenon but may fail to provide information on what the phenomenon is.

Informed by (Costantino, 2012; Lincoln and Guba, 1985; McCaslin, 2008; Mingers, 2006; Onwuegbuzie and Leech, 2005; Patomäki and Wight, 2000; Ratner, 2012a, 2012b; Smith, 2012)

2.3 Research approach

2.3.1 Quantitative, qualitative and mixed-methods approaches

The research approach refers to whether quantitative, qualitative or mixed methods will be used to collect data and answer research questions (see chapter 1 of Teddlie and Tashakkori, 2009 for explanation of three approaches). Although not definitive, research paradigms, ontologies and epistemologies often align with a particular approach (e.g.Teddlie & Tashakkori, 2009, p.88). As Cresswell (2003, table 1.4, p.19) suggests, both quantitative and qualitative approaches can align to the pragmatic paradigm. Using both approaches together is typically referred to as mixed methods research, and although it is a fairly new concept, the approach has been used for many years, although without using this specific terminology until more recently (Teddlie and Tashakkori, 2009). Perhaps because of that, definitions vary extensively, based on how the qualitative and quantitative approaches are used. The interpretation that will be used for this thesis, is that there are three types of mixed methods research: sequential, concurrent and transformative (Cresswell, 2003).

The sequential procedure allows for one approach to inform the other. The concurrent, or simultaneous, procedure allows for both quantitative and qualitative approaches to be used at the same time, to answer the same question perhaps, and requires that the findings are integrated to answer the question fully. The transformative procedure starts with a theory and applies that theoretical lens to the study, which happens to include both quantitative and qualitative approaches and can include sequential or concurrent procedures also (Cresswell, 2003, p.16).

2.3.2 Inductive and deductive approaches

Induction can be defined as *“the process whereby a general law is established by accumulating particular instances”* (Crotty, 1998b, p.31-32). In other words, inductive reasoning starts with the data and builds conclusions and generalisations from the data itself – a ‘bottom-up’ approach. Deduction can be defined as involving *“arguing from the general (e.g., theory, conceptual framework) to the particular (e.g., data points)”* (Teddlie and Tashakkori, 2009, p.23). This is a ‘top down’ approach, whereby a prediction, theory or hypothesis is stated, and then tested by collecting data.

2.3.3 Thesis research approach

This thesis aims to explore the quality of apps both in terms of quantitatively measurable criteria such as use of theoretical underpinnings and evidence-based content, as well as qualitative based criteria such as characteristics of apps and user perceptions of acceptability from feedback and user reviewers. To generalise, the truth (data) that is sought is not thought to be a single truth, but consists of multiple truths (perspectives) depicting the same phenomena (quality).Therefore both quantitative and qualitative approaches will be used. Three characteristics of mixed methods approaches should be explained: sequence, priority and relationship of the alternative approaches, otherwise known as method of integration (Denscombe, 2010). Cresswell et al. (2003) states that integration can occur at any or at multiple stages: during research question generation, data collection, analysis or interpretation. Additionally, Greene et al. (1989) report five purposes for using mixed methods, expanding on the definition of a complementary study by stating that *“qualitative and quantitative methods are used to measure overlapping but also different facets of a phenomenon…the logic of convergence requires that the different methods assess the same conceptual phenomenon.”* (p.258) (see table 2).

Table 2 Purposes of performing a mixed methods study.

Name	Purpose
Triangulation	Seeks convergence, corroboration, correspondence of results from the different methods
Complementarity/ Complementary	Seeks elaboration, enhancement, illustration, clarification of the results from one method with the results from the other method.
Development	Seeks to use the results from one method to help develop or inform the other method, where development is broadly construed to include sampling and implementation as well as measurement decisions.
Initiation	Seeks the discovery of paradox and contradictions, new perspectives of frameworks, the recasting of questions or results from one method with questions or results from the other method.
Expansion	Seeks to extend the breadth and range of inquiry by using different methods for different inquiry components.

Adapted from table 1 in (Greene et al., 1989)

Informed by these sources, a concurrent, complementary mixed methods approach was chosen. Neither the qualitative nor quantitative approach is dominant. Integration occurs during generation of the research questions and data collection in order to inform the overall findings and conclusions of the programme. Triangulation, although not a specific purpose, also occurs. Data analysis is in parallel and does not involve integration as quantitative data has not been transformed into qualitative data and the same is true of the reverse. However, a narrative mixed research synthesis using a segregated design is implemented to configure the complementary findings into the final argument and recommendations for future work (Sandelowski et al., 2006).

The mixed methods approach lends itself to use of both inductive and deductive reasoning and both will be used independently to answer the research questions. Qualitative data is explored and has inductive reasoning applied to generate a theory regarding features that are viewed positively or negatively by users. Using a deductive approach, quantitative data and descriptive qualitative data are captured to add weight to the existing theory that they contribute to the quality of apps or PA interventions.

2.3.4 Strengths and weaknesses of Mixed Methods

The main strengths of the mixed methods approach are its flexibility and its potential to reduce the impact of any inherent biases related to use of a single approach (Cresswell, 2003, p.15; Denscombe, 2010). These are particularly relevant when using both qualitative and quantitative approaches to answer the same research questions – as is the case here.

Mixed methods allows for both exploratory and confirmatory research (Denscombe, 2010; Teddlie and Tashakkori, 2009). For example, this thesis will confirm whether or not acknowledged quality indicators such as evidence-based content is present in apps, as well as explore whether other potential indicators of quality exist such as those relating to user perceptions. Similarly, it allows collection of a *'greater assortment of divergent views'* (Teddlie and Tashakkori, 2009, p. 33). The use of different approaches in this thesis could lead to a convergence of data on key quality components. For example, user comments could indicate that evidence-based content is something they perceive as indicating a high quality app. Alternatively, data could, and seems more likely to, diverge as the two different approaches have been used to ensure capture of a range of interpretations and perspectives on quality indicators. Essentially, using mixed methods removes the potential criticism that quantitative methods are only identifying established or accepted indicators of quality, or that qualitative methods are neglecting them and instead focusing on more in-depth issues of acceptability or usability at the expense of content, which is difficult to capture quantitatively. See table 3 for examples of strengths and weaknesses of qualitative and quantitative approaches.

Like both quantitative and qualitative approaches, the mixed methods approach also generates its own set of unique challenges (Denscombe, 2010). Time and cost are often cited as a barrier. However, for the present thesis neither present a challenge due to the nature of the work being carried out – for example, a common source of delay can be participant recruitment and qualitative analysis of lengthy transcripts, neither of which are required. In addition, apart from the cost of app download, no extra costs will be incurred from conducting a mixed methods design, as downloads would be required for both quantitative as well as qualitative appraisals. The researcher must have both quantitative and qualitative skills for mixed methods research. Again, this does not present a challenge as the researcher has experience of using both approaches. This also satisfies the next known issue, the tendency to oversimplify the distinction between qualitative and quantitative approaches. For example, while these are sometimes seen as mutually exclusive and evolving from certain philosophical stances, they may be less separate than first thought - qualitative data could include counts of words or phrases – this is known as content analysis in qualitative analysis, and quantitative questionnaires may include free text questions (Gray, 2018, p.194). The fourth weakness as described in Denscombe, (2010) is the supposed inability of the mixed methods approach to allow for emergent research designs, i.e. for methods and approaches to change based on findings generated during the study. However, Teddlie and Tashakkori (2009) argue that the mixed methods approach is so flexible and diverse that an adequate typology of mixed methods design cannot be created, and this is in part because it can include qualitative approaches with *"emergent strategies"* (p.139). In fact, Denscombe, (2010) goes on to acknowledge that there is "*no reason in principle why a mixed methods approach cannot adopt an emergent design*" (p.151). The final suggested weakness that is worth mentioning is the idea that different approaches may generate findings that don't corroborate each other. However, as mentioned above, the current thesis is not necessarily seeking the same findings from qualitative and quantitative approaches, therefore this is not a weakness that needs managing.

Although it has been specified above that inductive approaches will be used with qualitative data and deductive approaches with quantitative data, it can be argued that quantitative data interpretation will also have an element of inductive logic. Quantitative data, it has been said, cannot prove a theory, merely provide probabilistic evidence for it (Teddlie and Tashakkori, 2009, p.67-68). Therefore, to an extent, this is still discovering a pattern in data, as inductive approaches claim to do. It may be a more prevalent pattern perhaps, but it does not necessarily confirm the pattern is exhaustive.

However, deductive approaches have started to move towards a causal model of explanation (Teddlie and Tashakkori, 2009, p.71-72). Now they attempt to explain a phenomenon by determining the causal mechanisms that bring it about, rather than being based on an application of *'general laws'*. This makes their claims relatively easier to test rigorously, resulting in the ability to accumulate probabilistic evidence and therefore reinforce their claims. As such, quantitative data and deductive interpretations are more generalisable to other instances and populations.

The causal model of explanation is comprised of two key viewpoints: the idea that causation can be evidenced by statistical associations between a factor and an outcome, when all other factors are controlled for (regularity theory of causation) and the idea that a difference should be seen between those receiving the causal factor and those that don't (the counterfactual approach). In this thesis, both of these viewpoints can only be partially applied due to the non-experimental nature of the research. While it is hypothesised that apps will be more effective and more acceptable if they are of higher quality, only within group comparisons will be available during analysis, rather than comparisons to a group of control apps. Statistical analysis of associations between facets of quality and effectiveness and acceptability will also not be available due to the nature of the data (no raw effectiveness data is being collected) and the sample sizes, which are too small for a sufficiently powered calculation.

Table 3 Examples of the strengths and weaknesses of qualitative and quantitative methods

Approach	Strengths/Benefits	Weaknesses/Disadvantages
Quantitative	Facilitates rational decision making via identification of cause and effect relationships. Methods of quantitative data collection often include stringent ways of removing or reducing risk of bias so findings are valid and reliable. Results typically aim to be generalisable.	Methods of data collection can involve randomisation and controlling for bias can be difficult/impossible/ inappropriate/unethical in the real-world. Lacking ecological/external validity - much research is lab-based, so findings may not be replicable in the real-world. Often experimental/quasi-experimental designs fail to produce a definitive answer, despite this being what they claim to do – results from similar studies are often mixed. Ever-changing social world may be too complex. Explains what happens and how many times, but not why.
Qualitative	Aims to understand the perspectives of others rather than impose researcher's perspective/bias/theory. Social and content validity of data is thought to be high. Often results in in-depth, highly detailed data. Can prevent misrepresentation of a community (e.g. via ethnography). Can look for relationships, usually between themes, explain them and generate theories.	Can be costly in time and resources, for example if a specialist interviewer is required and lengthy interviews are conducted. Can be challenging for the researcher to remain engaged and gain understanding over the long term. Can be difficult to access participants depending on the nature and sensitivity of the research. Explains why something happens, but not how many times or what happens. Lack of generalisability.

	Can acknowledge contextual and historical influences – doesn’t operate in a controlled vacuum. Can be more accessible to certain population groups as a variety of data collection methods are available and can be used in different ways to generate data e.g. interviews, surveys, photos etc. Can access group processes and interactions.	Could be perceived as intrusive by participants. Recruitment can be difficult for small populations due to fear of identification and therefore ethical concerns may be more serious.

Informed by (Donmoyer, 2012; Lewis-Beck et al., 2004; Lincoln and Guba, 1985; Onwuegbuzie and Leech, 2005; Schensul, 2018)

The weakness of inductive reasoning is an exacerbation of the weakness of deductive reasoning – inductive interpretations can never prove a theory, but give a single instance of a pattern of data in a specific setting, with specific participants. Also, qualitative data and inductive interpretations are far less generalisable than quantitative outcomes. Qualitative researchers tend to use a different term: transferability – the readers' judgement (rather than the researcher's analysis based on sample size, statistical analysis and so on) of whether or not inferences from the research context can be applied to other contexts. Where a result is generalisable, it tends to be broadly applicable to other contexts. However, transferability may be more limited to personal or discrete experiences or exhibit certain degrees of transferability such as temporal, theoretical, population or ecological (Teddlie and Tashakkori, 2009, p.311-312).

2.4 Methodology

Teddlie and Tashakkori (2009, p. 21) define methodology as "*a broad approach to scientific inquiry specifying how research questions should be asked and answered. This includes worldview considerations, general preferences for designs, sampling logic, data collection and analytical strategies, guidelines for making inferences, and the criteria for assessing and improving quality*." They also state that "*methods are determined by…methodological orientation*".

The current thesis is informed by a pragmatic paradigm or worldview, which allows the research questions to dictate the most appropriate methods. It has the following research questions:

Research questions:

1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps that include feedback on immediate affect, including both apps developed for/by researchers and publicly available apps (commerical apps) in

the app stores?

3. What is the quality of these physical activity apps that provide feedback on immediate affect?

The thesis consists of two studies. Both address all research questions, following the concurrent mixed methods approach and segregated narrative mixed research synthesis design. However, in line with the pragmatic paradigm, different methodological choices for data collection were made due to the different sources of data. Individual choices for the two studies are explained below.

2.4.1 Methodological choice for study 1: Systematic review of peer-reviewed literature

With the rise in evidence-based medicine and practice (Guyatt et al., 1992; Olsen et al., 2009) and the consistent financial burden experienced by the UK NHS, currently due in part to an ageing population and increasing rates of obesity and multimorbidity (Stoye, 2018), there is now an expectation that health promotion activities will be chosen based on a demonstrable set of supporting 'proof'. However, determining the presence of a high quality evidence-base can be difficult. Mulrow (1994, p. 1) describes the landscape of study outputs as '*an enormous puzzle'* and advocates systematic reviews as a way to *'disentangle'* the '*puzzle's intricacies'.* The purposes of a systematic review are well established: keeping up to date with current research, identifying new and more effective interventions, discovering current interventions are ineffectual, collating guidelines for treatment/care, identifying more efficient ways of doing things, or delivering care and avoiding duplication and wasted research resources (Mulrow, 1994). The aims of this thesis align with some of these purposes. Seeking to identify any evidence of a high quality (and therefore potentially effective) new intervention (a PA app that provides feedback on affect), or any gaps in the evidence to direct limited resources towards future research.

Given the novelty of apps for behaviour change, as well as the recent rise in recognition of affect as being influential in behaviour, the literature was likely to be limited, requiring a review methodology that would comprehensively identify any relevant papers. According to a typology of reviews (Grant and Booth, 2009), this goal limited the potential type of review to six: meta-analysis, a state-of-the-art review, mixed methods review, systematic review, systematic search and review and systematised review. Meta-analyses only include quantitative studies. These studies are unlikely to be prevalent given the tendency to begin intervention development with qualitative exploration of participant needs or barriers and facilitators to use, and the newness of the field. The state-of-the-art review focuses on more current areas of interest and is time-limited in scope. However the extensive focus on quality for this thesis and the often lack of formal assessment within this sort of review, has led to the judgement that it was not an appropriate choice. Although it is acknowledged that the method of quality assessment performed in this thesis was different to those typically used in literature reviews. The mixed methods review typically includes a literature review, along with a form of qualitative study such as interviews or a combined qualitative and quantitative study literature review, to unite perspectives. The current thesis does aim to unite both quantitative and qualitative findings across the two studies, but mixed methods reviews that combine literature with an empirical study do not appear to be widely acknowledged, therefore this terminology was not used for the present thesis. The systematised review was also discounted as it was not considered a comprehensive collection of existing evidence. This left the systematic search and review, and the systematic review. The former typically addresses broad questions, more suitable to an assessment of the state of the literature across disciplines and domains, to identify gaps for areas of study. However, the newness of the app literature and the resurgence in the affect-for-behaviour change hypothesis, means that the area of study can already be identified, but what is required is a better understanding of the precise quality and gaps in that literature base. As such, a systematic review methodology was chosen.

The Centre for Reviews and Dissemination summarises the aim and function of a systematic review as *“to identify, evaluate and summarise the findings of all relevant individual studies, thereby making the available evidence more* *accessible to decision- makers"* (Centre for Reviews and Dissemination, 2009, p. v). This includes identifying gaps in the literature. Although there are acknowledged weaknesses of this methodology, such as the tendency to restrict eligible data by study design and focusing on effectiveness (Grant and Booth, 2009), these need not be adhered to while still maintaining the rigorous and replicable design.

2.4.2 Methodological choice for study 2: Systematic evaluation of public apps

In order to be comprehensive and gain as true a picture as possible of the app landscape and its potential strengths and weaknesses, public apps must be included. However, unlike literature, there are few established methods for exhaustively identifying, collating, evaluating and synthesising public apps. Previous studies of public apps have used principles of systematic review methodology to address this gap (Anderson et al., 2016; Knight et al., 2015; Reynoldson et al., 2014). However, all features of the systematic review process may not be suitable. For example, use of search terms to identify relevant apps is unreliable, as the app store algorithms are proprietary and therefore not publicly available to determine how searches function. Despite constraints, the strengths of the systematic review method – transparency, replicability, rigour, synthesis of all available knowledge, bias reduction – offer viable, defensible principles by which public apps can be identified and evaluated. While resources are limited and other methods undeveloped, such principles were applied in this thesis, termed a systematic evaluation in this instance.

2.5 Data collection methods

The data collection methods for the two studies are the same, because both ask the same questions of the apps, despite the apps being from different sources. Therefore, the methods discussed are relevant to both studies.

Both qualitative and quantitative data were being collected and a number of methods and tools were chosen to collect the characteristics and quality of the apps. This is because of the variety and lack of consensus in the literature for the best way to evaluate health apps (Jake-Schoffman et al., 2017). The data dictated the best method of collection.

Quantitative data is typically numeric and tools for data collection are often predetermined or have close-ended questions. Surveys, questionnaires, measuring instruments and experiments are often used to generate statistical data. In contrast qualitative data is typically text such as field notes or memos, verbalisations such as interviews, or images and documents. Tools for data collection may not be predetermined, but are developed during data collection. Open-ended questions are often used and grounded theory, case studies or ethnography are then used to generate and develop themes from the data (Cresswell, 2003, 2007; Gray, 2018). See table 4 for what quantitative and qualitative data in this thesis consists of.

Table 4 Quantitative and qualitative data collected as part of thesis

Quantitative data to be collected	Method of data collection	Qualitative data to be collected (including some descriptive data)	Method of data collection
Number of apps with feedback on affect as part of physical activity promotion	Data extraction form - count	App administrative characteristics	Mobile App Rating Scale (Stoyanov et al., 2015)
Number of apps that include credible development teams, and users in design of the app	Data extraction form – count	Characteristics of target population	Data extraction form - summary
Number of apps with theoretical underpinnings	Data extraction form – count facilitated by Theory Coding Scheme (Michie and Prestwich, 2010)	How feedback on affect is collected, processed, provided and for how long	Data extraction form – description
Number of apps where quality was formally assessed	Data extraction form – count	Characteristics of development team and process	Data extraction form – how users were involved, expert status of developers Mobile App Rating Scale - Affiliations, Developers
Whether or not quality issues were addressed	Data extraction form – yes/no	How quality was assessed	Data extraction form – description and name of tool used if applicable
Number of behaviour change techniques used	Data extraction form – count facilitated by Behaviour Change Taxonomy v1 (Michie et al., 2013)	Types of theoretical underpinnings of app	Theory Coding Scheme - descriptive list (Michie and Prestwich, 2010)
Number of behaviour change	Data extraction form – count	Type of behaviour change	Data extraction form –

techniques that are associated with physical activity improvement used	facilitated by (Gardner et al., 2015; Howlett et al., 2018)	techniques used	descriptive list facilitated by Behaviour Change Taxonomy v1 (Michie et al., 2013)
MARS quality assessment scores (Aesthetics, Engagement, Credibility, Information)	Mobile App Rating Scale – overall score and score per quality area	Type of behaviour change techniques that are associated with physical activity improvement used	Data extraction form – descriptive list facilitated by (Gardner et al., 2015; Howlett et al., 2018)
Effectiveness/Potential impact	Mobile App Rating Scale – score for perceived impact combined with evidence-base item score	Method of physical activity measurement	Data extraction form – description and name of tool used if applicable
Number of apps that include public health guidelines	Data extraction form – count	Reliability and validity of physical activity measurement	Data extraction form – summary of literature on tool
Rates of usage if reported	Data extraction form – percentage or however depicted in paper	Quality of apps as described by users	Data extraction form – descriptive summary of user reviews and thematic analysis of user reviews from stores
Number of apps with security/privacy features	Data extraction form - count	Currency and maintenance of the app	Data extraction form and Mobile App Rating Scale – recency and frequency of updates
		Types of security/privacy features	Data extraction form – summary
		Types of PA targeted by apps	Data extraction form – summary
		Description of app and technical	Data extraction form –

		aspects	summary Mobile App Rating Scale – Technical aspects

PA = physical activity

Pre-existing standardised data collection tools are preferable, as comparisons can be made between the outcomes of studies that use them. However, standardised tools may not always exist for the outcome of interest, or may not be entirely appropriate and need adaptation. Quantitative data collection was facilitated by using a combination of pre-existing and new tools, as data is predominantly a count of the presence or absence of certain features (see second column in table 4). Pre-existing tools used for quantitative data collection included the Theory Coding Scheme (TCS) (Michie and Prestwich, 2010), version one of the Behaviour Change Technique (BCT) Taxonomy (Michie et al., 2013), and the Mobile App Rating Scale (Stoyanov et al., 2015). A list of BCTs associated with changes in PA levels were used to determine the presence of such BCTs in the apps. The BCTs were identified in two review papers (Gardner et al., 2015; Howlett et al., 2018). The remaining quantitative items were collected using a standardised data collection form designed for this purpose for each study.

Qualitative data collection tools are often designed for purpose such as interview schedules, questionnaires and coding frameworks. This is due to the exploratory nature of qualitative research. However, for the purposes of this thesis, qualitative data has been sub-divided into two categories: descriptives and perspectives. Both relate to the evaluative nature of the methodology – describing what the characteristics of the app are and if the users' needs are met. Characteristics were captured using pre-existing tools including the TCS, BCT taxonomy and the PA BCT list to determine the theoretical basis of the apps and the types of BCTs used. Free-text items in a data collection form were used to capture additional quality criteria (e.g. user reviews and perspectives on quality indicators), some of which were informed by existing literature (see method of PA measurement and reliability and validity of that measurement below). Perspectives were also captured by collecting app store star ratings.

Both pre-existing and fit-for-purpose instruments were used to ensure all data of interest was captured (which pre-existing tools can neglect). Where possible, reliable or validated tools were used to ensure a high standard of research practice and allow for present and future comparisons of app quality. Both sets of tools ensure that only relevant data was captured, reducing wasted resources. However capturing user perspectives allowed for the potential for new quality indicators to be identified.

2.5.1 Theory Coding Scheme

The TCS is a framework for identifying theoretical components in interventions. It consists of 19 closed items that capture use of theories and their constructs, how they have been used to inform, tailor or develop an intervention, the degree to which the intervention content is linked to a theory/construct, and if theoretical components are measured, reported and discussed (Michie and Prestwich, 2010). There are descriptions of each item and a comprehensive set of guidelines and training materials (see appendix 1 for framework and guidelines).

The behaviour-change literature and its practitioners generally consider theory-based interventions to be preferable and more effective (McEwan et al., 2018). However, researchers and developers tend to use theory sparingly, inconsistently, don't use it at all, or report its use poorly and as such, there is mixed evidence for the use of theory-based interventions (McEwan et al., 2018; Prestwich et al., 2015). Therefore, it is important to identify theoretical components and report them appropriately to allow for further analysis of their effectiveness.

2.5.2 Behaviour Change Technique (BCT) Taxonomy

The BCT taxonomy is used to identify the presence or absence of features of interventions. It consists of 93 techniques, divided into 16 categories (Michie et al., 2013). It includes descriptions and examples of each BCT and training is also available online. The taxonomy facilitates reporting of interventions to ensure identification of effective techniques, easy replication and faithful implementation. It was developed using a Delphi exercise involving 19 experts in behaviour change, from around the world, from both research and practitioner backgrounds, as well as receiving feedback from an international advisory board.

The taxonomy, or earlier unofficial versions of it, have been used to identify components in interventions including apps (Bardus et al., 2016; Bondaronek et al., 2018; Free et al., 2013; Roberts et al., 2017), setting a precedent for its use in this thesis.

2.5.3 Behaviour change techniques associated with physical activity (PA) change

In order to try and address the potential quality issue of including BCTs that may not be associated with changing PA behaviour, a fit-for-purpose checklist was created based on two recent reviews (Gardner et al., 2015; Howlett et al., 2018).

2.5.4 Mobile App Rating Scale (MARS)

The MARS is an instrument used to determine the quality of an app in relation to its engagement, functionality, aesthetics and information (Stoyanov et al., 2015). Some or similar items are assessed during usability assessment and general user testing when developing new technology. Others have been considered part of assessments in recent app reviews. Previously (see Background chapter), it was established that engagement is a part of ensuring app usage and is starting to be investigated in the literature on digital interventions (Perski et al., 2017). Thus these MARS items can be considered important for quality (Brooke, 1996; Milward et al., 2016; Nielsen, 1993, chapter 5; Reynoldson et al., 2014). MARS allows users to make a subjective judgement of the app (Stoyanov et al., 2015) and captures basic descriptives of the app. Most questions have five possible answers and a non-applicable response, each with a unique description. A mean score can be generated for each section and an overall mean score across the first four quality sections can also be calculated (see appendix 2 for MARS).

2.5.5 Data collection/extraction form

The remaining quantitative and qualitative items were collected using a standardised data extraction form devised for each study. Additional items were chosen based on their relevance to the research questions, relevance to the behaviour domain, whether or not they had previously been considered relevant to quality in other app reviews, best practice for intervention development, and whether or not and how quality was assessed and addressed (see chapters 3 and 4 and appendix 3 for more details of the data being collected). The apps, where available, were downloaded to assess all characteristics of interest, in line with typical usability testing procedures (Nielsen, 1993).

To support data extraction for the systematic review and allow for better reporting and easier replication, the Template for Intervention Description and Replication (TIDieR) checklist informed the data extraction form (Hoffmann et al., 2014). Reporting and descriptions of interventions are often cited as challenging for reviews resulting in tools such as TIDieR, PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) (Moher et al., 2009) and the Quality and Risk of Bias Checklist for Studies That Review Smartphone Applications (BinDhim et al., 2014). All were used in this thesis.

2.6 Reliability and validity

Reliability and validity relate to how truthful and accurate results are. In particular, they can be ascribed to the tools or instruments or methods of data collection that are employed.

Reliability is considered to be whether or not an instrument will repeatedly produce the same results when measuring the same phenomenon – either over time, with nothing else changing, or between different users. Validity is '*whether one can draw meaningful and useful inferences from scores on the instrument*' (Cresswell, 2003, p. 157), in effect, whether or not the tool measures what it should be measuring. See table 5 for a brief description of each of the three main types of reliability and validity.

Table 5 Types of reliability and validity

Reliability or Validity	Type	Description
Reliability	Internal consistency	*“Are the items’ responses consistent across constructs?”*
	Test-retest	*“Are scores stable over time when the instrument is administered a second time?”*
	Consistency	*“Were errors caused by carelessness in administration and scoring?”*
Validity	Content	*“Do the items measure the content they were intended to measure?”*
	Predictive/Concurrent	*“Do scores predict a criterion measure? Do results correlate with other results?”*
	Construct	*“Do items measure hypothetical constructs or concepts? Do scores serve a useful purpose/have positive consequences when used? Have the right definitions been used to measure the variable?”*

Adapted from (Cresswell, 2003), p. 157-158, 171.

Reporting the reliability of instruments is important for rigorous research and replication, as well as to justify use of a tool (Cresswell, 2003). Where available, inter-rater reliability and internal consistency has been reported for the pre-existing tools that will be used in this thesis (see Methods chapter 3.4.7.4), demonstrating their quality. Their inclusion of extensive guidelines and/or descriptions or training materials should maintain consistent administration. In addition, a subset of data was checked by a second researcher to ensure correct interpretation and test-retest reliability (see Methods 3.4.9, 3.4.10, 4.4.7, 4.4.8 and Results 5.2.1).

The validity of the pre-existing instruments is less clear. Most are simple enough to assume content and construct validity, and for the purposes of this thesis also have construct validity in that they seek to identify characteristics of interest to inform future research. However, predictive validity is uncertain. Although tools and items have been chosen based on an existing understanding that the items they measure are synonymous with quality and effectiveness, the latter will need to be judged once results are collated. Threats to validity can also be internal or external, relating to inadequate procedures, contamination between groups, incorrect generalisations and inadequately powered statistical tests (Cresswell, 2003, p.171). Given the nature of the studies included, such threats are unlikely to feature except for the potential for incorrect generalisations, which will be examined closely in the Discussion chapters.

A more in-depth discussion of the validity and reliability of the tools and their outcomes will take place within sections 3.4.7 and 4.4.5 and the Discussion chapter.

Qualitative data considers reliability and validity differently to quantitative research and reliability is considered to play less of a role (Cresswell, 2003, p.195). Validity is considered to consist of concepts such as 'trustworthiness' 'credibility' and 'authenticity' (Cresswell, 2003, p.196). Alternatively, trustworthiness has also been defined by internal and external validity, reliability and objectivity (Lincoln and Guba, 1985). The use of different terminology for qualitative data is due to the nature of the data being collected and the tools that are used to do so. Participants, rather than instruments, are providing the truth, therefore the reliability and validity of the interview questions may be of less importance than the perspectives they generate and the context and personal view from which the researcher and participant are operating. As Lincoln and Guba's work over the years has been seen as a 'touchstone' for many interpretivists who seek to justify their findings, they will be used as the main resource to discuss these terms (Schwandt et al., 2007). These authors consider trustworthiness an overarching construct parallel to rigour, consisting of credibility instead of internal validity, transferability instead of external validity or generalisation, dependability instead of reliability and confirmability instead of objectivity (Schwandt et al., 2007). The main weakness of these terms is that they aim to parallel positivist or quantitative constructs, when qualitative research, as established, provides a different perspective. Therefore, the term authenticity has been used to establish the accuracy of qualitative interpretations (Schwandt et al., 2007). Authenticity is proposed to consist of fairness, ontological authentication, educative authentication, catalytic authentication and tactical authenticity (see table 6 for definitions of trustworthiness and authenticity constructs).

Table 6 Definitions of trustworthiness and authenticity

Construct and sub construct	Definition
Trustworthiness	
Credibility (parallel to internal validity)	*"whether or not the a research report is 'credible' to the participants whom the researchers studied"* Whether or not findings are *"credible to the constructors of the original multiple realities"*
Transferability (parallel to external validity or generalisability)	*" the transferability of inferences from particular sending context (the research setting) to a particular receiving context (other similar settings)"*
Dependability (parallel to reliability)	*"the extent to which the process of inquiry is dependable; the ability of the human instrument to yield consistent results"*
Confirmability (parallel to objectivity)	*"the extent to which the product of the inquiry is confirmable, including whether results are grounded in data, whether inferences are logical, whether there is inquirer bias, and so forth"*
Authenticity	
Fairness	*"it must be the case that different constructions will emerge from persons and groups with differing value systems…Fairness may be defined as a balanced view that presents all constructions and the values that undergird them"*
Ontological authentication	*"improvement in the individual's (and group's) conscious experiencing of the world…[it] ought to be to raise consciousness, or to unite divided consciousness, likely via some dialectical process, so that a person or persons (not to exclude the evaluator) can achieve a more sophisticated and enriched construction"*
Educative authentication	*"It is not enough that the actors in some contexts achieve, individually, more sophisticated or mature constructions, or those that are more ontologically authentic. It is also essential that they come to appreciate (apprehend, discern, understand)—not necessarily like or agree with—the constructions that are made by others and to understand how those constructions are rooted in the different value systems of those others"* *"increased understanding of (including possibly a*

	sharing, or sympathy with) the whats and whys of various expressed constructions"
Catalytic authentication	*"Inquiry, and evaluations in particular, must also facilitate and stimulate action"*
Tactical authenticity	*"whether the evaluation is empowering or impoverishing, and to whom"*

Informed by (Lincoln and Guba, 1985, p.296; Teddlie and Tashakkori, 2009, p.26, 296; Schwandt et al., 2007)

The qualitative data collected as part of this thesis is predominantly descriptive. Therefore, these findings are only subject to certain facets of trustworthiness– perhaps better captured within their quantitative parallels, such a reliability or objectivity, while authenticity may be irrelevant. For example, credibility seems irrelevant for the descriptive qualitative outcomes and tools, however dependability or reliability of the instruments is pertinent. Confirmability or objectivity is also relevant, and as such, attempts have been made to be comprehensive in the type and interpretations of quality within the digital and intervention literature, hence the use of multiple instruments both pre-existing and purposively developed. As also stated within the quantitative measures, a subset of qualitative data was double checked by another researcher to limit inquirer bias.

The user's perspectives - reviews or comments on the apps - can be subject to threats to both trustworthiness and authenticity however, and are better suited to this sort of assessment. It is possible that credibility may be threatened in study 2, the systematic evaluation of commercial apps, as only a proportion of the reviewers' comments, which are typically brief rather than rich, will be extracted for pragmatic reasons. In study 1, any comments captured in the literature will be extracted, but confirmation of their credibility (member checking) cannot be performed and data collection is restricted to what the papers report. However, credibility could be assessed by making comparisons between the results from the two studies, allowing a contrast between issues identified in a controlled context (as described in the literature) and free-living adults using apps in the real-world. Credibility of the user reviews in study 2 will also be considered during testing of the apps. Transferability could also be assessed by comparing results between studies, known as triangulation, which is a recommended method for ensuring accuracy of findings (Cresswell, 2003, p.196). Such triangulation should prompt reporting of discrepant information, another recommended activity (Cresswell, 2003). However, transferability in study 2 should be unthreatened as the data is being *'sent'* from a real-world context and will be transferred to a real-world context.

Descriptives and user perspectives may encounter threats to authenticity. This is because it is the researcher that will initially benefit in terms of ontological authentication and educative authentication. In terms of fairness, attempts are being made to capture a range of current interpretations of quality, as well as a range of perspectives from users. However in study 2 the sample size will be limited and comments are likely to be brief and cannot be explored further with the user. In terms of catalytic authentication and tactical authenticity, it can only be stated that the findings are hypothesised to inform future research to improve PA, to the benefit of individuals, society and the health system. It is for the reader to decide the extent of the impact of the present research on future action and the degree to which it empowers future individuals, but that is the proposed goal of the present (and indeed, of most) health research. Further discussion of the implications of the results of this work will feature in the Discussion chapter.

2.7 Chapter summary

As reported, a pragmatic paradigm was adhered to including perspectives from both critical realism and relativism, and objective and subjective viewpoints. A mixed methods methodology was chosen based on the research questions, comprised of a systematic review (study 1) and a systematic evaluation (study 2) (see table 7 for summary). Data collection methods include both quantitative and qualitative assessments using pre-existing and fit-for-purpose tools embedded into a data extraction form for each study. Various threats to reliability and validity of the findings do exist due to the nature of the data. Pre-existing tools and methods were chosen to minimise these threats where possible. In addition, the combination of methods allowed by the pragmatic paradigm should minimise bias and strengthen findings by way of triangulation and/or by seeking to be comprehensive by including both a breadth of acknowledged indicators of quality and allowing possible identification of new indicators through participant perspectives.

Table 7 Summary of methodology

Paradigm	Ontology	Epistemology	Methodology	Methods
Pragmatism	Pragmatism: Critical Realism and Relativism perspectives	Pragmatism: Objective and Subjective perspectives	Concurrent, complementary mixed methods, with mixed research synthesis using segregated design. Study 1: Systematic review Study 2: Systematic evaluation Both use quantitative and qualitative analysis	Quantitative assessment of quality TCS, BCT taxonomy, PA BCT checklist, MARS and additionally devised items. Qualitative description and assessment of quality using TCS, BCT taxonomy, PA BCT checklist, MARS, user reviews and additionally devised items.

BCT = Behaviour Change Technique; MARS = Mobile App Rating Scale; PA = Physical Activity; TCS = Theory Coding Scheme

3. Methods: Study 1: Systematic review of peer reviewed literature of physical activity apps with feedback on affect

3.1 Introduction to chapter

Until now it is unclear the extent to which affect has been used as a motivator for PA in app-based interventions in free living contexts, although there is evidence to suggest that meal-based positive affect is being explored as a way to promote normal eating (SMARTFOOD Emotion Trial (Renner et al., n.d.)). In addition, the literature cites many challenges for digital behaviour change approaches, with establishing quality being just one. These include conducting comprehensive, well considered and granular evaluations in terms of their effectiveness and usage, engagement (which is likely to be unique compared to more traditional interventions, is often cited as a problem, and is hypothesised to be due to apps not being user-friendly, respectful of privacy, trustworthy or useful), mechanisms of effect, cost-effectiveness and adherence to regulations and ethical standards (Michie et al., 2017; Miller et al., 2018; Torous et al., 2018). Given these challenges, app development and content is likely to vary, and with it, quality. For example, common sense dictates that development teams which include academics are likely to produce theory-based content, often interpreted as an indicator of good quality by an academic audience. In contrast, teams with experienced developers may have apps with better user interfaces. As such, there is already recognition that digital intervention development should be multidisciplinary (Michie et al., 2017; Murray et al., 2016). However, it's still unclear what the quality of the apps of interest might be.

A systematic review will be conducted to determine the evidence base from published literature. Following this, an evaluation of public apps available in app stores will be conducted (Chapter 4).

This chapter will describe the methods of a systematic review to identify literature-based PA apps that provide feedback on immediate affect and their characteristics and quality. The definition of quality will be discussed.

3.2 Defining quality

3.2.1 Compiling quality indicators

Scientific studies have established many potential sources of bias that influence perceptions of quality (Higgins et al., 2016; Shea et al., 2007; Sirriyeh et al., 2012). Unlike studies, apps for behaviour change are relatively new and therefore app-specific quality indicators have not been universally adopted, as demonstrated by the fact that studies of apps capture and call different features ‘quality indicators’. There are a handful of conducted and ongoing reviews that compile these indicators, showing variability in what has been captured and how (Billiet and Vanden Bûssche, 2016; BinDhim et al., 2014; Nouri et al., 2018; Van Velthoven et al., 2018). While Nouri et al., (2018) collate a variety of quality indicators into classes, they don’t solve how to define and capture each criterion (and indeed, cite the reported tools as having varied, or non-existent, criterion definitions). They also don’t include papers that haven’t used a formal tool or formalised set of criteria and include a number of tools that were developed for specific diseases or disorders (see Nouri et al., 2018). This meant that many criteria were related to disease-relevant content and recommendations which are not relevant for PA promotion. While reviewing the criteria across a range of studies, it also became clear that when definitions were provided, they were inconsistent, with usage and engagement often used interchangeably. The criterion of engagement is further complicated by a recent study that tries to conceptualise it as multi-faceted, consisting of both a subjective experience as well as a measurable behaviour synonymous with usage (Perski et al., 2017). Therefore it was considered necessary to determine a relevant set of criteria and definitions for the purposes of this thesis.

Subsequently, a scope of the more recent literature discussing and reviewing apps and user preferences for apps was performed. Using methods similar to Nouri et al., (2018) quality indicators were consolidated. Examples of often-mentioned key quality indicators – explicitly specified by authors in some instances, but not necessarily all – are listed and broadly defined in table 8 and appendix 4. While compiling this list, criteria were also included if they seemed pertinent to PA apps (see item 3), explaining why some have fewer references. Many of these indicators are based on traditional perceptions of intervention quality that have been established in behaviour change literature such as theoretical underpinnings, evidence-based content, intervention fidelity and so on. Others appear to be based on usability engineering, interface design and user testing heuristics developed for product and technology testing (Nielsen, 1993, 1994). The tools cited were not all created explicitly for apps; some were adapted from website quality assessment criteria and others were for assessment of systems in general (e.g. Brooke, 1996). While beyond the time-frame for the current thesis, future validation of these terms within the academic and app developer community would be valuable.

Table 8 Frequently reported indicators of app quality

Quality indicator		Summary definition	Examples of app assessment studies where indicator is captured without using a developed tool	Examples of tools using indicator or studies using the tool (NB: not all assessing PA apps)
1	Acceptability or Participant perceptions	Positive and negative feedback or recommendations from users on the app content e.g. preferences or recommendations for information, tone or features. Can include user ratings or reviews of the app (the latter may overlap with other indicators such as 4, or 10. These should be coded separately). May include barriers and facilitators (feasibility) to use of the app and/or smartphone such as it being easy to fit self-report requests into your routine, or forgetting to carry the phone. May also include whether or not users/providers/practitioners would recommend the app to others/patients	(Bondaronek et al., 2018; Dunton et al., 2011; Milward et al., 2016; West et al., 2012)	(BinDhim et al., 2014; Brooke, 1996; Murray et al., 2016; Powell et al., 2016; Stoyanov et al., 2015; Vasa et al., 2012)
2	Aesthetics	Visual attractiveness of the app interface design in terms of colours, fonts, and layout. How professional the design is. How pleasing to eye the design and layout is. Can include the relevance of design to the behaviour.	(Milward et al., 2016; Reynoldson et al., 2014)	(Jin and Kim, 2015; Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013; Stoyanov et al., 2015; Taki et al., 2015)
3	Behaviour measurement	The type of tool/method used by the app to	(Bort-Roig et al.,	(Martínez-Pérez et al.,

	tools (e.g. PA measurement)	measure physical activity (or other behaviours as applicable) and its related validity and reliability. For example, objective or subjective measures. Also can include the definitive accuracy of the PA tools for capturing physical activity or perceived accuracy of in-app calculations.	2014; Knight et al., 2015; Muntaner et al., 2015)	2013; McMillan et al., 2015)
4	Credible, trustworthy/appropriate or useful/essential information	Content of the app is likely to be accurate or believable – not making impossible or implausible claims. Content is safe for users, won't harm them or will minimise harm or provides a caveat for medical information that requires seeing a professional. Information/app appears useful.	(Peiris et al., 2014; Reynoldson et al., 2014; West et al., 2012)	(BinDhim et al., 2014; Jin and Kim, 2015; Loy et al., 2016; Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013; McMillan et al., 2015; McNiel and McArthur, 2016; Stoyanov et al., 2015)
5	Currency/Maintenance of the app and its documentation	The date of the last update/regularly updated (no consensus in the literature, range between 1 and 6 months since last update) and date of creation and last update are reported. App documentation is updated as well as the app itself.	(Bondaronek et al., 2018; Reynoldson et al., 2014)	(Jeon et al., 2014; Martínez-Pérez et al., 2013; McMillan et al., 2015; McNiel and McArthur, 2016; Stoyanov et al., 2015; Taki et al., 2015)
6	Development process and teams	Affiliations or credentials of app development team (University, Industry, Government, Commercial or Non-commercial etc.), involvement of experts and users in development.	(Bondaronek et al., 2018; Reynoldson et al., 2014)	(BinDhim et al., 2014; Jeon et al., 2014; McMillan et al., 2015; McNiel and McArthur, 2016; Powell et al., 2016; Taki et al., 2015)
7	Effectiveness/Potential	Evidence of improvements in relevant health	(Jake-Schoffman et	(McMillan et al., 2015;

	impact	outcomes/behaviours. Potential for impact on behaviour or health (can be measured by tools such as the Precede-Proceed Model which includes judgements of whether or not the intervention addresses predisposing, enabling or reinforcing factors that are thought to determine behaviour) (Has previously been interpreted as use of evidence-based content or theory (Bondaronek et al., 2018), but that is captured under items 9 and 12)	al., 2017; Peiris et al., 2014; West et al., 2012)	Powell et al., 2016; Stoyanov et al., 2015)
8	Engagement	Use of methods to encourage user interactivity with the app, can include use of certain strategies or features that promote/inhibit for example, feedback, tailoring, prompts/reminders, gamification. (Often this term has also been used to refer to usage/response to app intervention features or feasibility, such as required step-count submissions or required message responses (Monroe et al., 2015). Or it has encompassed both these and the above summary definition (Rose et al., 2017). These items have been separated out here into other criterion i.e. 13).	(Bort-Roig et al., 2014)	(Anderson et al., 2016; Cowan et al., 2013; McMillan et al., 2015; McNiel and McArthur, 2016; Powell et al., 2016; Stoyanov et al., 2015; West et al., 2013)
9	Evidence-based content/components	Use of techniques, strategies, information, practice or recommendations that are based	(Bardus et al., 2016; Breton et al., 2011;	(BinDhim et al., 2014; McMillan et al., 2015;

		on scientific evidence that demonstrates their usefulness. This can include behaviour change techniques or 'predictors' associated with improvements in behaviours in general or the target behaviour, prescribed behaviours or practices advocated by reliable government bodies such as Public Health England, or the National Institute of Health. Techniques can be captured by referring to taxonomies such as the Behaviour Change Taxonomy v1 (Michie et al., 2013)	Direito et al., 2014; Jake-Schoffman et al., 2017; Knight et al., 2015; Pagoto et al., 2013)	McNiel and McArthur, 2016)
10	Functionality/Usability	Ease of use of the app and/or smartphone features, such as navigation, terminology, design in relation to ease of use, not aesthetics (see 2) as well as general perception of how much support might be required for use or how complex or inconsistent it might be. Functional errors related to app operations such as bugs/crashing also captured here. Includes practicality of use for promoting or capturing physical activity based on functions and features. Can be assessed by questionnaires such as the System Usability Scale, interviews or user-testing/performance tests.	(Coughlin et al., 2016; Jake-Schoffman et al., 2017; Milward et al., 2016; Monroe et al., 2015; O'Reilly and Spruijt-Metz, 2013; Reynoldson et al., 2014)	(Anderson et al., 2016; BinDhim et al., 2014; Brooke, 1996; Jin and Kim, 2015; Loy et al., 2016; Martínez-Pérez et al., 2013; McMillan et al., 2015; McNiel and McArthur, 2016; Murray et al., 2016; Powell et al., 2016; Stoyanov et al., 2015)
11	Security/Privacy	Data privacy and/or security. For example, could include availability and accessibility of a privacy policy as well as its content, or a	(Bondaronek et al., 2018; Milward et al., 2016; Reynoldson et	(Anderson et al., 2016; Jin and Kim, 2015; Loy et al., 2016; Martínez-Pérez et

		required login for the app or ability to make personal content private rather than sharing with app community. Could include meeting Data Protection standards.	al., 2014)	al., 2013; McMillan et al., 2015)
12	Theoretical underpinning/components	Use of theoretical constructs from established behaviour change theories, or mention of use of theory, to inform development/content/evaluation of app in some way. Has/can be assessed by standardised tools such as the one developed by Doshi et al., (2003).	(Muntaner et al., 2015)	(Cowan et al., 2013; Roberts et al., 2017; West et al., 2013)
13	Usage/Compliance	Not to be confused with engagement, usage or compliance refers to responses to app content such a required step-count submissions or response to prompts to complete questions. Similar to fidelity.	(Bort-Roig et al., 2014; Monroe et al., 2015; Rose et al., 2017)	(McMillan et al., 2015; Murray et al., 2016)

PA = physical activity

3.2.2 Existing quality assessment tools

In the last few years, attempts have been made to develop quality assessment tools for health apps that include some of the reported indicators, but none are considered the gold standard. Other criteria have also been reported, but less frequently mentioned. (Martínez-Pérez et al., 2013; McMillan et al., 2015; Murray et al., 2016).

Other tools and methods of quality assessment or quality regulation also exist. For example Our Mobile Health (Our Mobile Health, n.d.) and ORCHA (ORCHA, n.d.) are both app evaluation websites. The latter uses a large number of criteria to assess apps in the app stores and gives them quality ratings. However, their full criteria are not in the public domain. The UK Medicines & Healthcare products Regulatory Agency suggests medical apps should incorporate a certification (CE) mark to ensure the app is *'fit for the purpose it claims and it is acceptably safe to use'* (Medicines & Healthcare Products Regulatory Agency, 2018, p.5), while the US Federal Trade Commission advises on relevant laws associated with development of medical apps (Federal Trade Commission, n.d.). The NHS Apps Library is currently beta testing assessment criteria (NHS Digital, 2018) and the British Standards Institute even specifies quality criteria specifically for health and wellness apps (British Standards Institution, 2015). Many of these quality criteria appear to overlap with those cited in table 8.

3.2.3 Challenges of developing quality assessment tools

Anderson et al., (2016) critiqued a number of existing quality assessment tools in the process of creating their own. The authors found similar quality themes among the tools, but limited subcomponents, and spent a long time creating a complex tool. Billiet and Vanden Bûssche, (2016) supported the complexity of the process by citing the need for skills including IT development, privacy and security issues, behaviour change strategies and medical knowledge just to develop a high quality app, let alone assess one. They also cited numerous issues with current assessment tools including the tendency to adapt from website quality assessment tools (Jeon et al., 2014;

Stoyanov et al., 2015), questionable generalisability (Stoyanov et al., 2015), lack of validation (Anderson et al., 2016; Loy et al., 2016), building a tool based on usability criteria alone (Anderson et al., 2016), and the failure to report the quality criteria at all (Yasini et al., 2016).

Despite the obvious weaknesses of existing tools, pragmatism and time-scales meant that a suitable development process for a new tool could not be carried out within the timeframe for the current thesis. Therefore, to ensure comprehensive assessment of key quality indicators, those reported in table 8 were captured. In order to capture as many criteria as possible, in a resource efficient method, an existing, tested quality assessment tool was used to allow for generalisability (Stoyanov et al., 2015), combined with an existing and tested comprehensive theory coding scheme (Michie and Prestwich, 2010) and taxonomy of behaviour change techniques (Michie et al., 2013) to ensure reliable and high quality assessments. Additional standalone items were included where the quality assessment tool did not already capture a key indicator. Further details on the tools and choice rationale, are in section 3.5.5.

3.3 Revised research questions and objectives

Based on the definitions of quality established in this chapter, the final research questions were adapted slightly.

Research questions:

1. Are there any physical activity apps in the literature that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps in the literature that include feedback on immediate affect?
3. What is the quality of these apps that provide feedback on immediate affect, where quality is defined as a multi-faceted concept consisting of 13 features?

Objective:

O1: A systematic literature review will be conducted to identify current evidence of physical activity apps that include immediate feedback on affect for adults. Quality and characteristics will be captured and assessed.

3.4 Review methods

The protocol for the systematic review was registered on PROSPERO, the international prospective register for systematic reviews, on 31st October 2018 (registration number: CRD42018107289, appendix 5).

3.4.1 Databases and search strategy

A range of databases were searched during scoping (AMED, ASSIA, BIOSIS, CINAHL, Clinical Trials, the Cochrane Library, EMBASE, HMIC, Index to Theses (ITT), OpenGrey, Psychology and Behaviour Sciences Collection, Science Citation Index Expanded, SPORTDiscus, Social Sciences Citation Index). The original aim was to conduct a highly sensitive search to identify relevant papers and use a machine learning tool called Abstrackr (Wallace et al., 2012) to facilitate automatic screening. This would allow a vast number of abstracts to be screened quickly, despite limited resources. However, on closer inspection of the software and following limited personal communications with the developer, it was difficult to determine exactly how Abstrackr functioned. Discussions with other users revealed that it was predominantly being used to reduce the number of screeners, rather than the number of abstracts that needed screening. This led to a revised approach to the literature review to ensure feasibility. As such, the databases and the search strategy were reduced to ensure a specific and manageable review was conducted.

Databases that returned few hits, exhibited persistent technical difficulties (ITT, ASSIA), became unavailable at the host institution (ITT, SPORTDiscus), had limited search functionality (BIOSIS, ClinicalTrials, OpenGrey, Science Citation Index Expanded) or returned hits which were likely to be captured elsewhere (OpenGrey), were removed from the strategy.

Databases were chosen due to their focus on medical, psychological and behavioural sciences. Scopus was included as it covered technology-based literature. Included databases also returned the largest number of hits when looking for literature on PA and affect in general. Databases were last searched 22nd March 2018 and included:

- CINAHL (EBSCO)
- EMBASE (Elsevier)
- MEDLINE (EBSCO)
- PsycINFO (EBSCO)
- Scopus (Elsevier)

The searches were limited to publications in the English language and from the year 2000 onwards. The date cut-off was chosen because according to one review, it appeared that PA measurement and promotion via smartphones began around the 2000s (Bort-Roig et al., 2014), and both the Apple App store and Google Play were opened in 2008. Apps were being created prior to 2000 but their popularity, prevalence and acknowledged opportunities for health were less established. Where applicable, type of publication was limited to academic journals, theses/dissertations, articles in press and published conference papers. Conference abstracts were excluded as it was likely that they had undergone a less rigorous peer review process. Although the emerging nature of the research area meant that conference abstracts may have revealed relevant apps, the limited information provided in such abstracts was deemed insufficient to enable quality assessment.

The search strategy was comprised of three sets of terms: physical activity, affect and smartphone or app terms. Medical Subject Headings (MeSH) terms and free-text terms were used with wildcard operators. Sets of terms were combined using Boolean operators. The search strategy for Medline is shown in figure 8 below. The strategy was then translated across the other databases, with minor differences occurring due to differences in database functionality (see appendix 6).

The strategy was developed based on scoping work. In addition, physical activity and smartphone terms were informed by existing PA and app literature reviews (e.g. Bort-Roig et al., 2014; Lamming et al., 2017), while affect terms were informed by a recent review of the predictive ability of affect for future exercise, a review of apps for mood tracking and preponderance for self-efficacy to be associated with affect either as a mediator or moderator or for PA, behaviour change, or as an outcome (Caldeira et al., 2017; McAuley et al., 2003; Pender, 2011; Rhodes and Kates, 2015).

Figure 8 Medline search strategy for systematic review

1. AB walk* OR TI walk*
2. (MH “Walking”)
3. AB exercise* OR TI exercise*
4. (MH “Exercise+”)
5. (AB “physical activity”) OR (TI “physical activity”)
6. (MH “Physical activity”)
7. (AB “physical* fit*”) OR (TI “physical* fit*)
8. (MH “Physical Fitness”)
9. (AB inactivit* OR inactive) OR (TI inactivit* OR inactive)
10. **1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9**
11. AB mood OR TI mood
12. (AB affect or affective*) OR (TI affect OR affective*)
13. (AB feeling OR feelings) OR (TI feeling OR feelings)
14. AB “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
15. TI “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
16. (AB emotion OR emotions OR emotional) OR (TI emotion OR emotions OR emotional)
17. (AB self-efficacy OR “self efficacy”) OR (TI self-efficacy OR “self efficacy”)
18. AB circumplex OR TI circumplex
19. **12 OR 13 OR 15 OR 16 OR 17 OR 18**
20. AB smartphone* OR TI smartphone*
21. (MH “Smartphone”)
22. (MH “Mobile applications”)
23. (AB “mobile phone*”) OR (TI “mobile phone*)
24. (AB apps OR app OR application*) OR (TI apps OR app OR application*)
25. (AB Iphone OR I-phone OR android OR iOS) OR (TI Iphone OR I-phone OR android OR iOS)
26. (AB “mobile health”) OR (TI “mobile health”)
27. (AB phone OR mobile OR telephone) OR (TI phone OR mobile OR telephone)
28. (AB “Mobile device*”) OR (TI “mobile device*”)
29. AB “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”
30. TI “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”
31. (AB “cell phone*” OR “cellular phone”) OR (TI “cell phone*” OR “cellular phone”)
32. (MH “Cell Phones+”)
33. AB digital OR TI digital
34. **20 OR 21 OR 22 OR 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31 OR 32 OR 33**
35. **10 AND 19 AND 34 (English)**
36. **Limiters – Date of Publication 20000101-20181231.**

3.4.2 Data management and deduplication

Database outputs were imported into Endnote X7.8 for data management and deduplication. Initial deduplication was performed by comparing the following paper details: author, year, title and ignoring spacing and punctuation. Secondary deduplication compared year, title and journal.

3.4.3 Inclusion/Exclusion criteria

Inclusion and exclusion criteria for the papers and apps are reported in table 9.

Table 9 Inclusion and exclusion criteria for systematic review

Characteristic	Criteria
Condition/domain studied	Physical activity promotion, not rehabilitation or physiotherapy or returning individuals to usual levels of functioning or improving balance. See 'Intervention' section below for further details.
Participants/Population	Free-living adults (18 years old or more). Apps could not be targeted at a specific clinical population (e.g. those with a specific disease, condition or disability or those recovering from one e.g. 'cancer survivors'), as specialist care is assumed. Overweight and obese adults were eligible and those who were 'at risk' of a disease or classified as pre-diabetic or pre-hypertensive. Active adults, but not professional athletes, were eligible for inclusion, despite the understanding that their motivation levels may differ to inactive adults who may also be more at risk from being sedentary.
Intervention	Intervention studies (where an app was used), development of interventions (apps), or evaluation of public apps, with the aim of increasing or exploring uptake, maintenance or promotion of, (lifestyle or structured) physical activity behaviour or decreasing sedentary behaviour/amount of inactivity were potentially eligible. Originally, apps focusing on multiple behaviours were going to be excluded, due to the complexity of determining content, features and quality that pertained only to physical activity promotion, aligning with other app reviews (West et al., 2013). However, during initial screening stages it became clear that, as expected, few eligible apps focused on physical activity alone. Therefore, it was decided that to ensure a comprehensive and informative review, apps were included if they targeted multiple behaviours including physical activity, and such apps would be discussed

	separately in the results. Apps that included, but weren't necessarily limited to, providing information about participant affect (positive or negative), preferably with a focus on physical activity-contingent affect, were included. Affect could be collected either immediately or within the same day, during or post-exercise, as evidence suggested during-exercise affect is associated with future activity levels (Rhodes and Kates, 2015).
Affect	Affect referring to any change in non-clinical affect, including but not limited to (mental) wellbeing, non-clinical stress, occasional anxiety (e.g. in response to risk or uncertainty) or low mood, enjoyment, happiness, pleasant or unpleasant feelings (Russell et al., 1989; Yik et al., 2011), and excluded any clinical measures of recognised affective disorders such as clinical depression as they represent a special population (e.g. bipolar or major depressive disorder) or chronic anxiety (e.g. Generalised Anxiety Disorder or obsessive compulsive disorder). Health-related quality of life was also included as certain subdomains such as 'emotional function' were considered relevant. Where identifiable, studies that focused on physical wellbeing, not mental wellbeing, were excluded.
Feedback on affect	Feedback had to go beyond measuring or monitoring affect without providing the results for review, and preferably within the context of physical activity. For example, studies would be excluded if participants' affect was measured using the Feeling Scale during and after exercise but results failed to be fed back to them to consider in relation to the exercise they'd performed. If participants received their affect measures, either interpreted by researchers in the context of the exercise they'd performed (e.g. "*you felt happier when you ran faster*"), or were able to continually view such measures within their physical-activity promotion app (i.e. in the context of their physical activity), these would be eligible for inclusion. Information could be available to the participant after completion of a measure for any duration of time that appeared to allow for consideration and learning. For example, if self-reported affect data was available for 24 hours, or it was stored and viewable for the duration of the study period, that would be eligible. However, if results disappeared after a single viewing, the app was not eligible. Where an app was involved but feedback on affect was delivered by another source e.g. website, wearable, written feedback or by counselling, these papers were

	included due to the expected dearth of apps with this function. A previous review of the effectiveness of apps for behaviour change suggested that supplemental material, wearables, or information from another source may be beneficial or preferred for behaviour change interventions that include apps, providing a rationale for this choice (Middelweerd et al., 2015; Zhao et al., 2016).
Apps	Native mobile applications were eligible for inclusion, unlike web applications. Compared to native apps, web apps can be restricted to certain browsers, require an internet connection, fare unfavourably in terms of efficiency and graphics, and are unable to make use of smartphone hardware and software such as in-built accelerometers or pedometers (an often-used resource for physical activity apps).
Language	Only studies reported in English were considered.
Comparators or control	Any control group was eligible for inclusion, including having no control group, as effectiveness was not a subject of this thesis, and any study design including qualitative or a description of the development of an app was eligible.
Types of study	Any study design was potentially eligible. Evaluations of publicly available apps were also eligible, as they had the potential to provide more extensive information about the app compared to information provided in the app store. Reviews were excluded, but those that appeared relevant had their reference lists hand-searched for outstanding relevant app studies. Protocols of relevant interventions (apps) were also eligible if they described an existing app (originally excluded, some protocols were found to give sufficient detail of existing apps to warrant inclusion).

3.4.4 Outcomes

A number of primary outcomes were sought, related to characteristics and quality, these are reported by research question, in table 10.

Table 10 Systematic review outcomes

Research question	Outcome
Are there any physical activity apps in the literature that include feedback on immediate affect (mood) to facilitate behaviour change?	Number of apps

What are the characteristics and content of physical activity apps in the literature that include feedback on immediate affect?	App name, platform, type of PA targeted, description of app, target population, technical aspects. Type of feedback on affect – PA-contingent or not. How affect was captured, processed, and subsequently displayed and for how long. Number of apps that assessed quality. How quality was assessed. If quality issues were addressed and how.
What is the quality of these apps that provide feedback on immediate affect, where quality is defined as a multi-faceted concept consisting of 13 features?	Acceptability or participant perceptions – narrative summary of user comments or reviews and/or app star rating, the latter derived from the administrative section of the MARS. Aesthetics – score from Aesthetics section of MARS. Type, reliability and validity of PA measurement tools used in app – narrative summary of evidence for measurement tool. Credibility, trustworthiness, and/or usefulness of provided information – score from items 13-19 in Information section of MARS (in order to make best use of MARS and allow comparison to other studies, the ‘appropriateness’ sub-item in this criterion has been subsumed under Engagement as appropriateness is captured under the Engagement section of MARS (item 5). In addition, item 19 which deals with effectiveness is captured here, and in the Effectiveness criterion again, to make results comparable between studies.) App currency and maintenance including app-related documentation where available – version number, release date and last update from administrative section of MARS and an additionally added item (release date). Development process and team – developer and affiliations as captured by administrative section of MARS and a narrative summary of development team and their relevant credentials (physical activity/app/behaviour change expertise). Number of apps which involved potential users in development and at what stage and how. Number of apps that assessed and/or addressed quality issues and how.

	Effectiveness/Potential impact (at the level of face validity) – score from item 19 in Information section and score from Perceived Impact section of MARS. Engagement – score from Engagement section of MARS. Evidence-based content and components – number of apps that use, and type of, evidence based-content/components, both behaviour change techniques as defined by the BCT Taxonomy and national physical activity guidelines for adults. Number of apps that use behaviour change techniques explicitly associated with change in physical activity levels, and type of techniques. Functionality or usability – score from Functionality section of MARS. Security and privacy – narrative summary of features. Theoretical underpinnings and components – Number of apps with theoretical underpinnings, theories used and how from TCS. Usage and compliance – rates of usage of app, any available. Any quality outcomes relating to these criteria reported in the paper.

BCT = Behaviour Change Technique, MARS = Mobile App Rating Scale, PA = Physical Activity, TCS = Theory Coding Scheme

3.4.5 Quality assessment of apps

As the literature suggests that there is no agreed upon gold standard of assessing quality of apps to date (e.g. BinDhim et al., 2014; Jeon et al., 2014; Michie et al., 2017), and those that do exist typically have limitations, a number of measures were used to capture the 13 different elements of quality and to try to reduce the emphasis placed on a single dimension or tool. The apps themselves were assessed by a primary researcher (LL) in the following ways, with a sample assessed by two researchers and disagreements were discussed with a third researcher as necessary.

1) Acceptability (1), Aesthetics (2), Credibility (4), Currency (5), Development team and process (6), Effectiveness (7), Engagement (8), Functionality (10)

The Mobile App Rating Scale (MARS) was used to assess quality indicators 1, 2, 4, 5, 6, 7, 8, and 10 (Stoyanov et al., 2015). The literature cites challenges such as the number of tools it's based on and the fact it's based on an adaptation of web assessment tools, (Billiet and Vanden Bûssche, 2016). However, it was chosen as it allows assessment of a range of key quality indicators, assesses them in a standardised and objective way, whereas other methods have shown poor inter-rater reliability (Powell et al., 2016) and was developed on the basis of evidence from multidisciplinary literature including the user experience literature for technology in general. It's also one of the few quality assessment tools that has been tested and validated and despite its limitations, is frequently used and considered acceptable to coders, even being endorsed by an Australian government health agency (Anderson et al., 2016; Billiet and Vanden Bûssche, 2016). Therefore, it facilitates comparisons between past and future app assessments. MARS was adapted to remove repetitive or redundant items. For example, minor changes to the administrative section reflected the fact that theoretical components, strategies and a brief description of the app were captured more thoroughly elsewhere. The search for an evidence base for the app was also performed before the quality assessment began to allow double extraction to be carried out smoothly. Where apps targeted multiple behaviours, MARS was adapted to ask coders to specify if their answers were definitely relevant to PA promotion or other behaviours. Where apps were unavailable (which was the case for all literature-based apps), MARS scores were assigned based on the information provided in the paper and a brief explanation of the reason for the score was provided to allow easier comparisons between coders.

In addition to MARS, indicator 5, 'Currency' will also be captured by an additional item designed for purpose that asks for the release data of the app. This item was added based on the tool developed by Martínez-Pérez et al., (2013) and how it assessed currency. Last update (from MARS) was combined with the new item, release date. It was considered to be the most straight forward method of assessing currency and likely maintenance of the app compared to other tools.

In addition to MARS, indicator 6, ‘Development team and process’, was assessed more extensively with free text items designed for purpose. Team credentials and expert status were evaluated as multidisciplinary teams are advocated (Billiet and Vanden Bûssche, 2016). Studies were also reviewed for if and how target users were involved in development and at what stage (e.g. during creation of app brief, testing prototype, testing final version etc.). This is considered relevant when developing or evaluating apps and eHealth technologies (e.g. Fougerouse et al., 2017; McMillan et al., 2015; Torous et al., 2018; Van Gemert-Pijnen et al., 2011). It is also established practice for intervention development in general, as demonstrated by the MRC guidelines on developing and evaluating complex interventions (Craig et al., 2008). These two items were informed by Bondaronek et al., (2018) and McNiel and McArthur, (2016) who had previously captured them.

2) Behaviour measurement tools – type, reliability and validity of PA measurement tools (3)

Type and quality of tool used to measure physical activity (indicator 3) was assessed using a free text item designed for purpose. Questions called for the tools to be identified and any strengths or weaknesses reported in the paper to be captured. Previous reviews suggest that the reliability of such tools in apps needs evaluating (e.g. Knight et al., 2015) and it has been proposed that when users perceive a measure to have poor reliability it reduces motivation to be active (Bickmore et al., 2009). Type of tool was categorised using the lists provided in Taylor (2014, p.48-55) and Strath et al., (2013). Most smartphones have inbuilt accelerometers and some have inbuilt pedometers. However, it was not feasible to determine the type of phone an app would be downloaded onto once out of the test environment, therefore an evaluation at the level of the smartphone was not performed (although limited evidence suggests Android accelerometers may have comparable validity to Actigraph monitors (Hekler et al., 2015)). Measures were subsequently evaluated based on the provided information as well as generally acknowledged strengths and weaknesses of objective and self-report measures (e.g. Taylor 2014, p.57; Sylvia et al., 2015).

3) Theoretical underpinnings (12)

The Theory Coding Scheme or TCS (Michie and Prestwich, 2010) was used to assess quality indicator 12 as the corresponding item in MARS was extremely limited. Unlike the 100-item scheme developed by Doshi et al., (2003) that has previously been adapted and used in app reviews (Cowan et al., 2013; West et al., 2013) the TCS has a manageable 19 items which assess explicit use of theory. The TCS also allowed for a range of theories and theoretical components to be identified – previous work suggests there are approximately 83 behaviour change theories (Michie et al., 2014). Although not used explicitly in previous reviews of apps alone, the TCS was used in reviews of behaviour change interventions that included apps and other mHealth modes of delivery such as Short Message Service - SMS (Garnett et al., 2018; Lyzwinski, 2014; Roberts et al., 2017). However, it has not specifically been adapted to code apps or digital interventions according to the second author, Andy Prestwich, (personal communication July 2018). Finally, another strength of this framework is its inclusion of predictors. Not only does it allow for interventions to be informed by theory or incorporate theoretical constructs, but it also accounts for additional, non-theory related predictors of behaviour change that have been acknowledged in the literature.

A previous review of digital interventions adapted the TCS, reducing it to 17 items as items 13 and 14 focused on methods, rather than theory and were deemed uninformative (Garnett et al., 2018). As the focus of the present review was to identify the quality of apps rather than the quality of app studies, items 12-19 of the TCS were not coded as they assess the use of theory in relation to evaluation of the intervention, focusing on the analysis and discussion sections of the papers.

There is mixed evidence with respect to the efficacy of theory-based interventions (Prestwich et al., 2015) and whether or not the theoretical basis is a factor behind intervention effect (Glanz and Bishop, 2010). However, the use of an underpinning theory is still widely considered a demonstration of quality, and of interest to researchers (West et al., 2013), as shown by the persistent production of reviews that seek to reveal and test the evidence of theory-based interventions (e.g. Arambepola et al., 2016; Cho et al., 2018).

4) Evidence-based content/components (9)

In order to assess quality indicator number 9, three approaches were taken.

Apps were assessed for the inclusion of BCTs in general using the BCTTv1 to determine whether or not they were present (Michie et al., 2013). Based on the taxonomy training, where BCTs were considered present they were given either a ‘+’ or a ‘++’ to indicate the strength of conviction that it was present. BCTs have been linked with efficacy and quality, and are often assessed when determining content, efficacy and quality of apps (e.g. Bondaronek et al., 2018, McMillan et al., 2015, Middelweerd et al., 2014, Jake-Schoffman et al., 2017). BCTs were not coded when they did not appear to be promoting a behaviour, for example when providing information about antecedents related to a menstrual cycle (Clue).

Apps were also assessed for the use of BCTs that are shown to be associated with PA promotion. Identified BCTs in apps were compared to a list of promising BCTs for PA change derived from Gardner et al., (2015) and Howlett et al., (2018). As the former includes adults of any activity status, who are healthy or who have some mild health conditions (hypertension), while the latter focuses only on inactive healthy adults, different BCTs have been identified as being promising. Both papers were chosen to inform the list in an effort to be comprehensive, as this area of research is still under review. Twenty-six BCTs identified as being linked to changes (either increases or decreases) in PA in predominantly healthy adults were extracted. Where papers disagreed, a BCT was labelled as having mixed-evidence. Originally a numerical scoring scheme was proposed for these BCTs and is reflected in the data extraction guidelines, however, this was subsequently not used following data extraction as it was considered uninformative.

Finally the presence or absence of public health guidelines for PA for adults, using a dichotomous, yes/no question designed for purpose, was captured (for details see Department of Health and Social Care, 2019; World Health Organisation, 2010).

5) Security and privacy features (11)

Apps were assessed for their provision and consideration of security and privacy features using a free text item. Assessors were prompted to report any evidence such as information on privacy protection, security policies, import and expert practices, use of login passwords, encryption, cloud back up and adherence to Data Protection standards (informed by Anderson et al., 2016; Jin and Kim, 2015; Martínez-Pérez et al., 2013; McMillan et al., 2015).

6) Usage/Compliance (13)

Apps were assessed on whether or not usage or compliance data was captured, how it was captured and the findings, using free text.

In addition to MARS, any quality results including user reviews, or indicators reported in the paper were also captured using a free text item (results of quality assessment) separately to inform the evaluation. For indicator 1, ‘Acceptability’ this was important, as there is debate over whether or not star ratings are indicators of popularity rather than content quality (Billiet and Vanden Bûssche, 2016). However, there is also concern over the use of reviews as indicators of quality, due to the fact that a user’s understanding of the app topic may be limited, making it difficult to provide a reliable judgement, the inability to be sure of the veracity of reviews (some positive reviews are paid for by app publishers) and the likelihood of oversampling extremely positive or negative experiences (BinDhim et al., 2014; Jake-Schoffman et al., 2017).

3.4.6 Quality assessment of papers

As the primary focus of the review was on identifying apps and assessing their quality, a formal assessment of study quality related to design and reporting was not performed. However, the Quality and Risk of Bias Checklist for Studies That Review Smartphone Applications has been used to ensure comprehensive reporting of the methods of the review (BinDhim et al., 2014).

3.4.7 Bias, reliability and validity

There were a number of sources of bias within this study, but where feasible, efforts were made to reduce them.

3.4.7.1 Databases

Although a small number of databases were searched, those chosen returned the largest number of papers and included both medical, behavioural science and technology literatures. In addition, hand-searching references of relevant reviews allowed for any uncaptured papers to be included. As this is a new field of research however, it was anticipated that few eligible papers would be returned at all. Grey literature was not searched, but scoping of such databases suggested few papers existed and were of limited relevancy.

3.4.7.2 Search strategy

Search terms were derived from existing reviews of apps and PA interventions and a scope of the literature, therefore it's unlikely that relevant apps were not captured, unless poorly reported. However, older papers that used less standardised terminology in their titles and abstracts may have been missed. To accommodate this, the term 'telephone' was included in the search strategy to increase the sensitivity of the search. In addition, any technology-based interventions that were not clearly irrelevant, but not clearly an app, were retained during screening to allow further assessment of their relevance during full text screening.

3.4.7.3 Inclusion and exclusion criteria

As it was hypothesised that very few eligible apps were likely to exist that targeted PA alone, apps targeting multiple-behaviours were also eligible. This caused problems of attribution, with quality assessments potentially reporting on aspects that are relevant to promotion of another behaviour. Good practice is to determine the exact 'active ingredients' in an intervention that produce the desired outcome, as well as providing a detailed account of the content of an intervention. Such granular and un-confounding assessment of app effectiveness and components has yet to be achieved. Findings for apps targeting multiple-behaviours will be less reliable than those that focus only on PA. However, they should still provide important information on the characteristics of feedback on affect and demonstrate whether or not it can still be considered an untapped motivational technique in the PA and digital intervention domain. Therefore these apps are important in order to establish the full evidence base for the area of interest. This argument is also true of those apps that use additional materials and supplements to the app to capture affect and feed it back. However, apps targeting multiple behaviours will be discussed separately, with a caveat for the reliability and validity of their quality.

An inclusive approach was taken regarding the eligibility of different populations. Overweight and obese adults were eligible, as were those who were merely ‘at risk’ of a disease, or were classified as ‘pre-diabetic’ or ‘pre-hypertensive’. This was in order to be as inclusive as possible in defining a non-clinical population. However, it could be argued that such individuals could be seeking regular medical treatment which includes prescribed exercise programmes, including a prescribed PA app, making this classification inappropriate.

In addition, active adults (but not athletes) were eligible. They were included as it’s possible that individuals with different motivation levels may prefer more or less complex app interfaces, which may impact user perceptions (a quality indicator) (Burns et al., 2012). In addition, it was deemed important to know if all apps appeared to target already motivated, active individuals, suggesting a gap in the intervention literature.

3.4.7.4 Quality assessment tools

3.4.7.4.1 Mobile App Rating Scale

Other tools and methods for quality assessment of apps or digital interventions exist (Bondaronek et al., 2018; Muntaner et al., 2015; Peiris et al., 2014; Reynoldson et al., 2014). However, as yet there appears to be no consensus on a gold standard (e.g. Reynoldson et al., 2014). MARS has been used to assess apps (e.g. Bardus et al., 2016), and appears the most feasible and comprehensive instrument to capture the range of different and complex facets of quality. Evidence suggests strong internal consistency and inter-rater reliability for the items of MARS (Stoyanov et al., 2015). However, the item capturing evidence-base did not have its reliability tested in the validation study.

Although MARS attempts to capture theoretical components, this section was considered extremely limited and weak by the research student, hence the inclusion of the TCS instead. Other weaknesses include that MARS can only be fully completed if the app under assessment is available for download, and it’s unclear how to code the app when it targets multiple behaviours. It was considered likely that few apps described in academic papers would be available due to proprietary rights or difficulty in maintaining the app, for example once awarded funding ceased. However, attempts were made to access them. Therefore, for the purposes of apps that were not available for download or focused on multiple behaviours, their scores could be perceived as biased and therefore they are discussed separately in the results.

MARS also attempts to determine the evidence for an app, including whether it has evidence from one or multiple randomised control trials (RCT), which is aligned with the higher scores by the item. The idea of the RCT as the gold standard for determining effectiveness may not be appropriate for apps in the first instance. Instead, different study designs and intensive iterative testing should be performed before an RCT is conducted (Jake-Schoffman et al., 2017; Michie et al., 2017). Given the recency of this literature, it seems unlikely many RCTs will exist in this field. In addition, use of an RCT does not guarantee a high quality study, it only suggests that RCT standards of implementation should have been met, not that they were definitely met (e.g. Brainard et al., 2016; Montori et al., 2006), meaning that this item may be biased.

3.4.7.4.2 Theory Coding Scheme

The TCS is more comprehensive and flexible than other theory coding schemes as already discussed, and development of the framework indicated that it has good inter-rater reliability (Michie and Prestwich, 2010).

It is important to determine whether the theory being used is appropriate for the behaviour/setting/population in question. Unfortunately, the TCS does not include a determination of the relevance of the utilised theory/constructs/predictors for the target behaviour, which may bias the perception of theoretical quality.

3.4.7.4.3 Behaviour Change Technique Taxonomy

The taxonomy demonstrates sources of bias. Although strong inter-rater reliability scores were reported for the taxonomy, not all BCTs had reliabilities calculated (Michie et al., 2013). Secondly, while taxonomy training was completed by the primary coder to ensure reliable coding, it was complex and lengthy and a second trained coder was not available for all secondary extractions, weakening the reliability of the findings from this tool. However, this is still an improvement on other studies using the taxonomy that have not reported the use of trained coders (e.g. Direito et al., 2014). In addition, the training recognises the importance of identifying techniques in intervention descriptions, but it does not emphasise making a distinction between interventions that use techniques multiple times, compared to those that only use a technique once. Studies have been conducted to determine the effectiveness or correlations between using more or less BCTs and outcomes (e.g. McEwan et al., 2018), but as yet it appears that use of the same BCT multiple times has yet to be explored (although use of a variety of motivational tools for PA promotion has been advised (Bielik et al., 2012)). This is worth exploring in future, given the evidence base for some BCTs over others.

3.4.7.4.4 Behaviour change techniques associated with physical activity change

Like the TCS, the BCT taxonomy, while useful for identifying BCTs, does not provide information on their appropriateness for the target behaviour, or on whether combinations or numbers of BCTs are appropriate. A recent scoping review suggested that evaluating the effectiveness of individual or combinations of BCTs is particularly challenging and current methods are limited (Michie et al., 2018). Another review suggested that certain BCT clusters are more frequent in theory-based interventions than others, but that there is negligible difference in effectiveness between theory or non-theory-based interventions, except for certain BCT clusters that appear to rely on theory to exert effect, complicating matters further (McEwan et al., 2018).

Therefore, to address the potential quality issue of including BCTs that may not be associated with changing PA behaviour, a fit-for-purpose list was created based on two reviews (Gardner et al., 2015; Howlett et al., 2018). Two studies were selected as they were both recent, and therefore likely to include the most up-to-date literature. In addition, they focused on two slightly different elements of PA – one looked at promoting PA behaviour change and maintenance in inactive healthy adults (Howlett et al., 2018), while the other looked at reducing sitting time in adults (Gardner et al., 2015), making the selection of BCTs they reported on more comprehensive and inclusive. It's possible that this introduced positive bias to the identification of certain 'good' BCTs, as the apps focused on one behaviour (promoting all type of PA) more than the other (reducing sitting time in general).

3.4.7.4.5 Data collection/extraction form

The remaining quantitative and qualitative items were collected using a standardised data extraction form devised for the study. Despite being informed by previous work or tools where possible, these items were not validated and may have been subject to poor reliability and validity. However, attempts were made to ensure comprehension by producing and including specific definitions of quality indicators (table 8), piloting the form with an additional researcher familiar with the app literature, and use of extensive guidelines to facilitate standardised data extraction and assessment.

Finally, and perhaps most importantly, to date it does not appear that quality assessments using these tools have been performed with only descriptions of the apps as provided in peer-reviewed papers. Scoping and personal communications with academics suggest that apps developed and reported in the academic literature may not be available, due to concerns over intellectual property, inability to maintain the app, or versions discussed may no longer be available due to updates. Therefore this is a novel approach, and as such, where apps are unavailable, they are discussed separately as scores may be biased by limited information.

3.4.7.5 Quality assessment in general

A discussion of the literature on evaluating health apps suggests a number of challenges for the present quality assessment, specifically: establishing content when implementation and operationalisation is so varied and changes based on use, additional payments or updates (Jake-Schoffman et al., 2017). In each case, data could be unreliable. Therefore the following mitigating practices were implemented: apps were downloaded where possible and updates prevented to ensure comprehensive collection of content; the same devices were used and detailed, explicit guidelines for quality assessment were produced and piloted to ensure understanding and promote consistency between coders. Where apps were available they were tested for 48 hours to allow new features and content to emerge. However, as mentioned previously, anecdotal evidence from discussions with authors of papers that assessed individual apps suggest that few apps would be available for download, a weakness of both that literature and this systematic review.

3.4.8 Ethics and approval

As this study collected secondary data, that did not involve directly collecting data from participants, it was not subject to ethical review. This is in accordance with previous literature reviews of apps that have stated that ethics approval was not required (Mateo et al., 2015; Roberts et al., 2017). Participant data within the papers should have already been anonymised and participants should have been selected without prejudice and consented without coercion, for the purposes of the original study. In this respect, the data provided by original participants is synonymous with data from qualitative studies, which is also routinely re-collected and reported as part of qualitative literature reviews, without seeking consent from the original participants.

The topic of interest was considered low-risk and did not require collection of any sensitive information. Apps were downloaded for review onto project devices, rather than a personal device, therefore private data was not at risk at any time. Mock information was entered into the apps to test functionality. Real data were not required.

There are many issues around the ethics of digital research. Despite the age of the internet, it is only in recent years that ethical review boards have begun to understand the potential risks, and guidelines have begun to appear such as those produced by the Association of Internet Researchers (AOIR) (Markham, 2012). Scientific societies such as the British Sociological Society, British Psychological Society and British Society of Criminology, are starting to include digital research in their statements of ethical practice and align with principles provided by the AOIR (e.g. British Sociological Society, 2017). Previously, data collected from the internet was considered and processed in terms of ethical considerations in much the same way as text data such as newspaper reports (see discussion in Flick and Tiidenberg, 2018). Therefore it was considered to be unaffiliated with a person. However, newer recommendations now suggest using the human subjects model when making ethical considerations regarding digital data (Flick and Tiidenberg, 2018). This suggests that text-based data written by people in a digital context such as a web-forum, social media site (or subsequently an app), should be considered as linked to the person and able to cause them potential harm and/or identify them if not sufficiently protected and anonymised. Therefore traditional concepts such as those defined in the Declaration of Helsinki which outlined principles of ethical research now need to be adhered to. However, for the purposes of the systematic review, these ethical standards are irrelevant, as they should have all been adhered to as part of the original studies.

3.4.9 Screening

The systematic review was conducted following the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). Screening criteria and tools were developed for titles, abstracts and full texts. They consisted of Excel spreadsheets where coding decisions were captured, and a set of guidelines for screening (See appendices 7-9 for guidelines).

3.4.9.1 Title screening

Following deduplication, a conservative approach to title screening was taken to avoid loss of potentially relevant papers. Titles were screened on five criteria: obvious irrelevance, mention of an app or other allusion to technology that could include apps such as use of the terms ‘digital’ or ‘ecological momentary assessment’, whether it was a review, if it focused on PA and if the participants were free-living adults with no obvious serious health conditions. During piloting of approximately 100 titles, two additional items were removed from screening: whether it was a duplicate (this could be noted in a separate notes cell rather than requiring explicit coding for all titles) and whether feedback on affect was mentioned, as it was expected that this would likely feature in the abstract rather than title. Criteria were also clarified to make coding easier (for example participants were additionally referred to as ‘target group for change’ as some studies used parents to enact interventions for children, which caused some confusion for coding).

Following piloting, the primary researcher (LL) screened titles for inclusion. A random selection of 25% (n=1226) of titles were divided between two additional researchers and screened to check consistency. The selection list was generated using www.random.org and the Random Integer Set Generator. Screeners agreed on codings for approximately 83% of papers (where both the reasons for exclusion and the final decision to include or exclude were compared, and absolute agreement was sought), which extrapolated to approximately a 4% error rate for the total set of papers (n=4901). When only looking at discrepancies for the decision to include/exclude, percentage agreement was approximately 91% and chance-corrected inter-rater reliability calculated using unweighted Cohen’s Kappa = 0.31, $P < 0.0005$ (Cohen, 1960). This is considered only fair agreement (Landis and Koch, 1977). However, when the discrepancies were examined, it was found that poor agreement was due to the tendency for the first coder

(LL) to be very inclusive, therefore although the inter-rater reliability was poor, it was poor in a sensitive way. In fact LL had a sensitivity value of 75.6% (31/41 included studies) resulting in more potentially irrelevant papers being passed through to abstract screening (a Type I error) rather than eligible papers being incorrectly excluded at title screening (a Type II error). Therefore, the rate of agreement was considered acceptable and the need to second-screen the remaining 75% was rejected. Agreement statistics were calculated using the Statistical Package for Social Sciences (SPSS) v20.

3.4.9.2 Abstract screening

Missing abstracts were located where possible (n=5), otherwise papers were sent through to full text screening without having abstracts screened (n=2). Abstracts for eligible papers and protocols were screened based on three criteria: focus on PA promotion, reference to use of a native app and adult, free-living participants with no obvious serious health conditions. Piloting of the screening tool on 100 abstracts led to slight changes in wording of criteria and clarification of exclusions. For example, if the intervention involved standard text-messaging this could be excluded. However, if there was a suggestion that a text messaging app was being used or an app that sent messages, it was retained in order to clarify the functionality of the app.

Based on abstract screening performed by LL, it was agreed that few papers (n=1) were likely to be eligible following full text review. Therefore, it was agreed that one of the eligibility criteria would be expanded. The focus on PA behaviours to the exclusion of other behaviours was expanded to include papers that examined apps exploring other behaviours (e.g. nutrition behaviours) in addition to PA (as reported in the Inclusion/Exclusion criteria for the intervention). Titles which had been initially excluded due to this criterion were re-screened which resulted in an additional 44 papers being passed through to abstract screening. Abstract screening guidance was updated to reflect this change and abstracts that were originally excluded were re-screened. Each abstract was screened independently by a second researcher and discrepancies resolved by a third researcher as required (see appendix 8 for guidelines).

3.4.9.3 Full text screening

Twenty texts were scoped by LL to ensure relevance before submitting to full screening. These were comprised of abstracts that reported an undescribed ‘telephone-based intervention’, and unfamiliar technology-based interventions such as interactive voice response (the latter was subsequently considered not to be an app). These were scoped separately as it was deemed unlikely that they would include an app and the number of full texts for screening was extensive. Full text papers were screened independently by an additional researcher. Any outstanding discrepancies were resolved by discussion with a third researcher as required. Where papers provided insufficient detail to enable screening (e.g. in relation to the app and feedback) and if other linked references had been cited, these were viewed for clarification and/or authors were contacted. Reference lists of relevant looking reviews identified during screening (and excluded as ineligible documents) were hand-searched for potentially relevant additional studies which were also full-text screened if their titles did not immediately exclude them.

The full text screening tool checked papers for the inclusion of a focus on PA, reference to a native app, provision of feedback on participant affect and a healthy adult free-living population. During piloting an item was added so that all inappropriate papers could be screened out (editorials, reviews), but notes could still capture whether relevant linked papers should be accessed and screened. Guidelines accompanied the full text screening tool (Excel spreadsheet) and were adapted during piloting to allow for poor description of potential apps and feedback and to prompt further exploration via contact with the author or linked papers. Piloting by an additional researcher familiar with the app literature also resulted in ‘unsure’ coding categories being removed, further clarification of apps, feedback and affect, and papers being classed as only ‘include’ or ‘exclude’, without a ‘review’ option (see appendix 9 for guidelines).

3.4.10 Data extraction

Prior to data extraction, authors of the included papers were contacted, app stores were searched, and the app names were used to search the original five databases targeted by the review, to determine if the app was accessible for download and if there were any subsequent papers providing evidence on the app. Search terms included app names, both long versions and acronyms, combined with first author names and affiliations using Boolean operators (see appendix 10 for search strategies). Searches were limited by date to 2000 onwards. Results were sorted by relevance and the first 50 entries checked for eligibility using title and where necessary, abstract. Papers that appeared to assess the quality or effectiveness of the app alone were retained. Papers were not retained if apps were tested on ineligible populations as defined in section 3.4.3 or were combined with other intervention content without individual evaluation of the target app. The additional searches were limited due to time and resource constraints. A sensitive screening process was carried out by the primary screener. Double checking was not performed as the search terms and eligibility criteria were considered sufficiently specific and straight-forward and it was expected that key papers would and should have been highlighted by authors/developers. Eligible papers, reported trial registries, authors’ publications and Google Scholar citations of the original paper were also checked for additional citations describing the eligible app.

A previous review of apps has advocated downloading apps to assess their full content (West et al., 2013). Therefore, where available, the app was downloaded onto a compatible device (Samsung Galaxy S6 phone (Android 7.0) or iPhone 6 (iOS 12) and trialled for 48 hours to allow comprehensive extraction. If the app was available in the app store, a link to the app page was also collected to allow completion of MARS. Apps, store links and extra papers (where available), as well as the original paper comprised the full data-set for data extraction. Where the app and other evidence was not available/published, data extraction was performed on the original paper alone.

A standardised, piloted data extraction form was created and used independently by LL. It integrated both the data extraction and quality assessment of the apps (using the aforementioned tools and items) and was informed by the Template for Intervention Description and Replication (Hoffmann et al., 2014) and adapted for purpose (see appendix 3 for guidelines and table 11 for details).

Piloting of the data extraction tool which included all of the quality assessment items, for a single paper by two researchers led to the following adaptations to the form and guidelines, reflected in previous sections:

- Guidelines were adapted to better facilitate quality assessment without an available app, or for those apps targeting multiple behaviours.
- Assessors were asked to report any information provided on the strengths and weaknesses of the PA measurement tool, including references to relevant papers describing its evidence-base.
- Repetitive items such as theoretical underpinning/strategies in MARS were removed where other tools captured the features more comprehensively (TCS and the BCT taxonomy in this instance).
- Subjective Quality section of MARS was removed as it was deemed inappropriate for researchers, rather than users, to code these items. The subjective nature of the questions meant that double coding would be redundant as both coders' perspectives would be valid.

Two apps (Haptivity, Health Mashups) were independently extracted and assessed by two additional researchers, one of whom had completed training for coding intervention components using the Behaviour Change Technique (BCT) taxonomy (Michie et al., 2013). A third app's extracted data was double checked by an expert in health interventions (Motimate). These were purposively sampled as they represented the range of apps and paper types: promoting PA alone, promoting PA and other behaviours, and a protocol paper. Discrepancies were resolved by a third researcher as required.

Table 11 Extracted data for systematic review

Data type	Data extracted
Administrative details of paper	Authors, title, year, country study was conducted in, app availability and availability of extra evidence for app.
Methods	Study type, design, duration, whether or not quality assessment was conducted, how and if issues were addressed, whether or not usage was captured and how.
Participants	Number, setting, diagnostic criteria, age, sex, country, socio-demographics, ethnicity.
Interventions	Number of groups, specific intervention - with focus on details of app and any relevant associated tools/content – name, description, targeted behaviour(s), type of PA targeted, platform, developers' affiliations and credentials and involvement of users, type of affect captured and method and duration of capture and feedback, use of tailoring, technical aspects, theoretical underpinnings, presence of PA recommendations for adults, and if in an app store: cost, size, store rating, number of ratings and version, release date and last update; presence of behaviour change techniques and those associated with PA change, security and privacy features.
Results	Detailed overview of quality outcomes measured by authors, MARS, TCS.
Other	Miscellaneous comments and key conclusions of relevance to the review questions, references to other relevant studies or linked papers that describe reported apps further.

MARS = Mobile App Rating Scale, PA = Physical Activity, TCS = Theory Coding Scheme

3.4.11 Analysis and synthesis

Inter-rater reliability was calculated for MARS scores and BCT coding for each app using the SPSS v20 and STATA v14.1.

The study protocol published in PROSPERO stated that Krippendorffs alpha was to be used to assess inter-rater reliability due to its ability to accommodate various types of data, numbers of rater's and missing data (Hayes and Krippendorff, 2007; Krippendorff, 2011). However, Cohens Kappa (K) (Cohen, 1960) has been used more extensively to assess inter-rater reliability in relation to coding BCTs, and allows for a weighted version (Cohen, 1968), applicable to ordinal data, such as the Likert scale data from MARS. Data also met the required assumptions to proceed with a Cohens K (Cohen, 1960). As MARS coding used a linear Likert scale, where a score of 1 compared to a score of 5 indicated a higher level of disagreement than a score of 1 compared to a score of 2, it seemed appropriate to consider the data as ordinal, rather than categorical (nominal), and analyse it using a linear weighted statistic (as opposed to a quadratic weighting).

Cohens K has previously been criticised for allowing too low a score to be considered acceptable agreement (McHugh, 2012), therefore the revised interpretation of level of agreement, reported in the same paper was used. As there was a high rate of negative agreement for BCTs (where both coders agreed a BCT was absent), the Brennan and Prediger coefficient (Brennan and Prediger, 1981) (equivalent to prevalence-adjusted bias-adjusted kappa, or PABAK (Byrt et al., 1993)) was also calculated, as such agreement can sometimes result in erroneous, low agreement statistics despite high observed agreement (Cunningham, 2009). Percentage agreements were also calculated.

Discrepancies were discussed and resolved to allow for a single mean score and agreed set of BCTs to be reported.

Descriptive data were tabulated including frequencies, percentages, means, medians and ranges calculated as appropriate.

Findings were combined with those from the app evaluation (Chapter 4) tabulated and a narrative, mixed research synthesis using a segregated design, conducted (Popay, 2006; Sandelowski et al., 2006). The segregated mixed research synthesis was chosen as the thesis methodology proposes that the quantitative and qualitative findings will complement each other – each bringing a different facet of quality to light and therefore neither can be translated into the other by 'quantitising' or 'qualitising'. Also, the aim of the thesis is to configure these findings into an argument that allows for recommendations to be made, a goal of mixed research synthesis (Sandelowski et al., 2006). Categorical strategies were used to reduce data into data chunked categories relevant to the research questions (Teddlie and Tashakkori, 2009, p.25).

3.4.11.1 Analysis of subgroups or subsets

Results were divided and reported in discrete groups: public versus literature-based apps and apps that targeted multiple behaviours, compared to those that only targeted PA.

4. Methods: Study 2: Systematic evaluation of publicly available apps

4.1 Introduction to chapter

Despite their rigorous development and evidence-based content, literature-based apps represent a small proportion of those currently available to the public. In some cases literature-based apps may not be generally available to the public for a long time, or potentially never, and they may be more at risk of becoming outdated and obsolete due to the reliance on finite funding, resources and staff. As such, it would be misguided to ignore the voluminous and popular publicly available apps – the other part of the app landscape.

While apps that include researchers on the development team may result in evidence-based content, publicly available apps are likely to be advertised to users in a more attractive way (Freeman et al., 2017) as developers strive for downloads, and be more engaging to users as engagement can mean profits from adverts, subscriptions or in-app sales. Researchers and developers can learn from both types of apps. The following chapter outlines the methods for a systematic evaluation of public apps available in the Apple and Google Play app stores.

4.2 Defining quality

The same 13 quality criteria and definitions specified in the previous chapter (3.2) were used for the app evaluation.

4.3 Research questions and objective

Aim: To assess the characteristics and quality of public apps in the app stores that use feedback on immediate affect to promote physical activity.

Research questions:

1. Are there any publicly available physical activity apps in the app stores that include feedback on immediate affect (mood) to facilitate behaviour change?

2. What are the characteristics and content of publicly available physical activity apps in the app stores that include feedback on immediate affect?
3. What is the quality of these apps that provide feedback on immediate affect, where quality is defined as a multi-faceted concept consisting of 13 features?

Objective:

O2: A systematic evaluation of publicly available apps in the app stores will be conducted to identify apps that include immediate feedback on affect for adults. Quality and characteristics were captured and assessed.

4.4 Evaluation methods

4.4.1 Identifying apps

Multiple app stores and clearinghouses exist including stores for different smartphone operating systems such as the Apple App store, Google Play, Blackberry app world, Ovi by Nokia, Palm app catalogue, Windows Mobile marketplace and Amazon App store. In addition, other clearinghouse sites that focus on health apps for clinicians or patients also exist: iMedicalApps (iMedical Apps, n.d.), AppScript (IQVIA, n.d.), HealthTaps's (HealthTap, 2010) AppRx, NHS Apps Library (National Health Service, n.d.), and so on (Boudreaux et al., 2014). Closer inspection of the range of clearinghouses reveals challenges both from a user and researcher's perspective. Challenges include ease of access, for example some are no longer open to the public (Happtique), or provision of limited, out of date, app lists (iMedicalApps). Some require registering for an account (AppScript, HealthTap) or include few apps (AppScript, NHS Apps Library). The stores for the different operating systems also vary, but Android and iOS (Apple) smartphones have the biggest market share compared to other operating systems (Chau and Reith, 2019; Statista, 2017). Google Play and the Apple App store also hold more apps than the other stores (Dogtiew, 2019).

Therefore, the evaluation focused on the most prevalent smartphone platforms and their stores.

Due to the larger number of apps on, and downloaded from, Google Play compared to the Apple App store (Business of Apps, 2016) and the prevalence of Android smartphones (Holst, 2019), the evaluation of commercial apps focused on those for Android devices. In practice, this meant that where duplicates occurred between stores, the Android version was retained for evaluation.

4.4.1.1 App store access

App Annie (App Annie, 2010) was searched for relevant apps. App Annie required free registration to access the top charts and allowed for a more usable interface to access a large numbers of Android apps in a single point in time compared to Google Play. It also allowed access to past daily rankings, ensuring a more replicable data collection method.

4.4.1.2 Identification approach and rationale

The top 100 free and top 100 paid Android apps from Google Play and iOS iPhone (not iPad) apps from the Apple App store, in the ‘Health & Fitness’ category of each store, were sought. This category was chosen in line with a recent evaluation of PA apps by Bondaronek et al., (2018). (Originally AppBrain (AppTornado GmbH, 2010) was going to be used to identify Android apps, however duplicates were found within the Top App lists that suggested this site may be less reliable and/or change more frequently, making it difficult to get a static list for long enough to extract the relevant data.) The top free and paid charts, rather than top downloads or top overall rated app charts were chosen for the following reasons:

1) Scoping of the top downloaded and top rated apps charts on AppBrain indicated that the number of PA apps based on their titles in the top 200 from each category was roughly equivalent to the number in the top 200 free and top 200 paid apps (as of 19/20 of July 2018).

2) Correspondence with the author of a recent review of publicly available PA apps (Bondaronek, 2017; Bondaronek et al., 2018) identified a small number containing an affect-based component. They were released predominantly by the same developers and used the same method to capture affect. Of these, a large subset appeared between the 400 apps included in the top downloaded and top rated charts. When compared, a proportion were found to appear between the top free and paid apps, suggesting that known apps showcasing a particular method of affect capture are not necessarily being excluded by using top free and paid UK charts.

Both free and paid apps of any cost were considered, unlike previous reviews which restricted their evaluation of paid apps (West et al., 2013).

Apps in the UK (ranked by the store using their proprietary algorithms), with descriptions in English, were assessed for relevance. App charts can be accessed by country or in some cases, worldwide. The UK store charts were chosen for the following reasons:

1) Availability of apps in app stores can vary between countries (Google, 2019; Apple Inc, 2019). Therefore, worldwide rankings of apps may have included a large number of foreign language apps, reducing the number of potentially eligible apps.
2) App Annie reported that they didn't display worldwide lists as they were not available within the stores themselves (App Annie Support, 2018). Therefore where this is offered, for example by AppBrain, it's likely that another set of algorithms are being used, exacerbating the 'black box' problem of chart-list creation.
3) Worldwide top charts did not represent what the user saw when visiting Google Play. When users navigated to the Health & Fitness category, they would see 'Top Selling' and 'Top' app charts, among others. As Powell et al.(2016) states, users are likely to pick apps they see soonest, rather than continue through the list and pick a later app. Therefore, free and paid charts were considered the closest parallel to

the charts displayed in the app stores and therefore most representative of the user experience. However, it is worth reiterating that the algorithms used to derive these ‘top’ charts are proprietary and so it cannot be confirmed how these charts were derived, despite speculation that downloads, reviews and star ratings are being combined in some way (Bankhead, 2017; Bardus et al., 2016; BinDhim and Trevena, 2015).

4) Searching UK app stores ensures relevance to the UK setting and increases transferability of findings and recommendations to a UK context (Anderson et al., 2016)

4.4.2 Types of Apps

Eligibility criteria of apps were identical to that of the systematic review (see 3.4.3). A brief overview is provided in table 12 for ease of reading and additional criteria reported.

Table 12 Characteristics of eligible apps for app evaluation

Characteristic	Criteria
Users	Free-living adults (18 y.o+). Apps must not be targeted at a specific clinical population.
Target behaviour	Smartphone apps with the primary aim of increasing, (lifestyle or structured) PA behaviour or intentions or decreasing sedentary behaviour/amount of inactivity were searched for. Apps that focused on multiple lifestyle behaviours including PA were eligible for evaluation.
Wearables	Apps that connected to a wearable such as a Fitbit, were included, but wearables were not sourced to check full functionality due to cost restrictions.
Affect	Affect referred to any change in non-clinical affect including but not limited to (mental) wellbeing, non-clinical stress, occasional anxiety (e.g. in response to risk or uncertainty) or low mood, enjoyment, happiness, pleasant or unpleasant feelings, and excluded any clinical measures of recognised affective disorders such as clinical depression (e.g. bipolar or major depressive disorder) or chronic anxiety (e.g. Generalised Anxiety Disorder or obsessive compulsive disorder). It was considered unlikely that this would be specified in a public app; however when relevant, health-related quality of life was included as certain subdomains such as ‘emotional

	function' were thought to be relevant for this evaluation. Where identifiable, apps focused on physical wellbeing, not mental wellbeing, were excluded.
Feedback on affect	Apps that included feedback on user affect (positive or negative), preferably with a focus on PA-contingent affect, were eligible for evaluation. Feedback could be available to the user after completion of a measure for any duration of time that appeared to allow for awareness raising, consideration and learning. In addition, information could be provided during or directly after being active, or at any time during app ownership. A recent review of apps that collected affect suggested that there was a variety of methods of capture available, including custom words or a note, predefined words, colour-scales, emoji's, audio or pictures (Caldeira et al., 2017). Therefore a broad interpretation of methods of capture was allowed including these options. Capturing affect could be apparent or referred to in the app description or screenshots as an indicator of potential feedback on affect, prior to downloading the app. If apps had a basic and deluxe version that were listed separately within the app store top charts, they were considered as separate apps (e.g. Breton et al., 2011). If there was an in-app upgrade option, that was free to do and accessed more features, this was considered part of the same app.

4.4.3 Outcomes

Outcomes of interest were identical to those collected for the systematic review (3.4.4), with one change.

- Themes relating to quality of the apps as derived from a sample of the 50 most recent user reviews, extracted from the date the app was downloaded onwards.

4.4.4 Quality assessment of publicly available apps

Quality assessment of the publicly available apps was conducted in almost exactly the same way as for the literature-based apps. A proportion of apps were assessed by two researchers using the 13 quality indicators and methods of assessment reported previously in sections 3.4.5 and 3.4.10. Disagreements were discussed with a third researcher as necessary.

One change included how the Development team was assessed. Public apps were assumed to include technology experts where they were distribution by established companies. Websites were searched for other details of development teams such as behaviour change experts.

The Quality and Risk of Bias Checklist for Studies that Review Smartphone Applications was used to ensure reporting of methods of the app evaluation (BinDhim et al., 2014).

4.4.5 Bias, reliability and validity

4.4.5.1 App stores

It's possible that relevant apps were missed as only two app stores were searched. However, given that the two chosen were the largest and cater to the most popular device operating systems, this was considered a minimal risk.

4.4.5.2 Identification approach

Although search terms have been used in previous studies assessing publicly available PA apps (e.g. Knight et al., 2015; West et al., 2013), due to the proprietary nature of the app store algorithms it was unclear how apps might be chosen and returned following the search. This meant the identification process would be less replicable as algorithms could change over time. Using top charts, and a website that allowed access to back-dated charts, meant that replication would be more feasible and reliable.

In addition, there has been a proliferation of guidance for developers to enhance downloads, known as 'App Store Optimisation' (e.g. Zolotareva, 2017). This has included the importance of keywords – often identified by just looking at other high ranking apps (e.g. Lamattina, 2016). Therefore, it's possible that terms may not have been reliable indicators of content or purpose and were subject to bias in order to enhance downloads. Testing of PA terms in Google Play (conducted summer 2018) suggested that terms were not sufficiently discriminatory and returned apps with irrelevant descriptions. Testing in App Annie (autumn 2018) showed terms were similarly unreliable.

Therefore, fewer relevant apps may have been captured, biasing the findings from the review and resulting in an incomplete data set, but in this instance replicability and reliability of methods were considered preferential. Future strategies to identify relevant apps may require input from app store technicians.

4.4.5.3 Types of apps

Like the systematic review, the inclusion of apps targeting multiple behaviours may have caused problems of attribution, but the argument from study 1 stands (see section 3.4.9.2). They are acknowledged as problematic and will be discussed separately.

The limited information provided in app descriptions and screenshots suggests relevant apps may have been missed. Short of downloading and trialling the full set of 400 apps, there's little that could have been done to mitigate this issue. Pragmatically, this was not feasible in the time frame or in terms of monetary cost. However, to address the issue of missing potentially relevant apps, a revised screening approach was devised. Where app descriptions and screenshots did not mention affect, but otherwise met the eligibility criteria, they were retained and categorised as 'unsure'. If after downloading, apps that demonstrated feedback on affect in their description were found to be ineligible and a sample size of close to at least 10 apps could not be reached, this retained group was sampled randomly for download and eligibility checks (see screening section 4.4.7 below). This was a novel and feasible approach to managing poor reporting in app descriptions and the likelihood of low numbers of eligible apps. However, it is acknowledged as a potential weakness which may have biased findings.

4.4.5.4 Quality assessment tools

4.4.5.4.1 Mobile App Rating Scale and Theory Coding Scheme

While MARS was more easily and reliably completed as apps were available for download, the TCS and item 19 of MARS may have been subject to bias due to limited information. It was hypothesised that few public apps would have had their development (or effectiveness) reported in the academic literature and therefore information on theoretical underpinnings would likely be restricted. In order to address this, developers were contacted, app websites viewed, and five databases searched for relevant evidence (see 4.4.8).

4.4.5.4.2 User reviews

User reviews were extracted for additional information on user perceptions of the quality of the apps. Reviews were collected in addition to star ratings, because there has been debate over whether or not star ratings are indicators of popularity rather than content quality (Billiet and Vanden Bûssche, 2016). However, there has also been concern over the use of reviews as indicators of quality, due to the fact that a user's understanding of the app topic may be limited, making it difficult to provide a reliable judgement, the questionable veracity of reviews (some positive reviews are paid for by app publishers) and the likelihood of oversampling extremely positive or negative experiences (BinDhim et al., 2014; Jake-Schoffman et al., 2017). Having said that, any qualitative data collected from people is subject to bias and researchers have little opportunity to verify individual accounts, short of performing member checking (or respondent validation). This is challenging due to the dichotomy between an individual's view and a researcher's perspective of an entire dataset, as well as still being subject to members sharing *'a common myth'* or desire to mislead (Lincoln and Guba, 1985, p.315; Mays, 2002; Varpio et al., 2017). However, credibility of the reviews was established during testing of the apps by the research student by comparing her experience of the app with that of reviewers such as functionalitiy and errors. Additionally, the use of the most recent reviews should minimise the issue of reviews seeming unreliable, if in fact they report on an older version of the app. In addition, previous studies have found distinct differences between positive and negative reviews, suggesting that at least negative reviews can be informative and are likely to be of value for determining user perspective (Vasa et al., 2012). Although other evidence from industry suggests that positive reviews are of more value to developers, as reviews are considered trustworthy by app users (Rhodes, 2019). User reviews have previously been assessed to guide developers in future design of high quality apps and are seen as powerful tools for users to self-assess app quality (Khalid et al., 2014; Vasa et al., 2012).

4.4.5.5 Quality assessment in general

A discussion of the potential issues with establishing app quality has already been reported in section 3.4.7.5 for the systematic review (study 1) and applies to study 2.

4.4.6 Ethics and approval

Human data was collected in the form of user reviews posted to the app store websites for eligible apps. Therefore, unlike the systematic review, the ethics of this data collection needed further consideration. In accordance with the Association of Internet Researchers ethics recommendations (Ess and AoIR ethics working committee, 2011), protection of individual user privacy was considered based on the acknowledged public nature of the app stores. App stores and their user review functions are clearly in the public domain, as reviews can be accessed without any form of registration or login.

The nature of the data was not sensitive, constituting reviews of PA app functionality and preferences. In addition, app store guidance for reviewers asks that no personally identifiable information is posted within a review (e.g. Google, n.d.) However, user names or handles and any identifiable information were removed/anonymised before analysis and storage. Analysis resulted in the use of quotes, which were anonymised before reporting. In addition, apps targeting adults were sought. Therefore user reviews should not have included reviews from minors which may have be more problematic, due to safeguarding rules and ethical practices for research involving children.

Finally, the app store functionality meant that reviewers could not be contacted to gain consent to use their reviews. Therefore, as the reviews were considered public and reviewers could not be contacted, no consent process took place or was deemed necessary. However, as human data was being collected, approval was obtained from the Chair of the Humanities, Social and Health Sciences Research Ethics Panel at the University of Bradford (Ethics Checklist EC25643, 3rd April 2019, appendices 11,12).This precaution was taken despite being unable to identify any manuscripts analysing user reviews that also reported seeking ethical approval for exporting them from the app stores (e.g. Guzman and Maalej, 2014; Khalid et al., 2014; McIlroy et al., 2016; Vasa et al., 2012; Wiles et al., 2018).

4.4.7 Screening and deduplication

4.4.7.1 Phase 1 – App charts

The top 100 free and top 100 paid apps in the UK, in the Health & Fitness category for both the Google Play and Apple App stores, displayed in App Annie were extracted from their charts on 31st October 2018 using screenshots to capture time and date of extraction.

App titles and hyperlinks to the app descriptions were transferred to an Excel spreadsheet and duplicates were removed based on title, developer, logo, description, pictures and price, with Android apps taking precedence. During deduplication it was found that two paid apps from Google Play were listed as being free. As it was unclear why this had occurred (again due to proprietary algorithms of the stores and App Annie) an additional two apps were extracted from the end of the paid list to avoid bias (apps numbered 101 and 102 from the original top chart.)

4.4.7.2 Phase 2 – App titles and descriptions

A standardised, piloted, proforma was used to assess app titles, descriptions and screenshots where provided. They were assessed for whether they targeted PA, targeted a specialist/clinical group or non-adults, referred to capturing affect at all, or were in English (see appendix 13 for guidelines). Piloting resulted in the following changes to the inclusion/exclusion criteria:

- Apps were included if their main focus was to track or self-monitor PA, for example pedometer apps, heart rate or pulse monitoring apps. These were included as self-monitoring is considered a behaviour change technique and therefore constituted actively promoting PA.
- Comparatively, apps that only provided map routes for walking or cycling, timers for interval training or calculated weights for weight lifting training were excluded. These were not thought to represent a behaviour change technique and would likely support activities already being performed, rather than promote their performance.

Excluded apps included period trackers, sleep timers and pregnancy apps, among others. As it was expected that few apps would be eligible, an inclusive approach was taken to identifying apps with a PA component. Apps that targeted multiple behaviours (e.g. nutrition behaviours) in addition to PA were included, meaning that weight-loss apps were often included.

Where affect capture/feedback was not suggested, this criteria was coded as 'unsure'. These apps could then be sampled from if initially downloaded apps (phase 3), revealed too many ineligible apps to reach a preferred sample size of approximately 10 for evaluation.

Two researchers screened each of the apps for relevance and where necessary, outstanding discrepancies were reviewed by a third researcher.

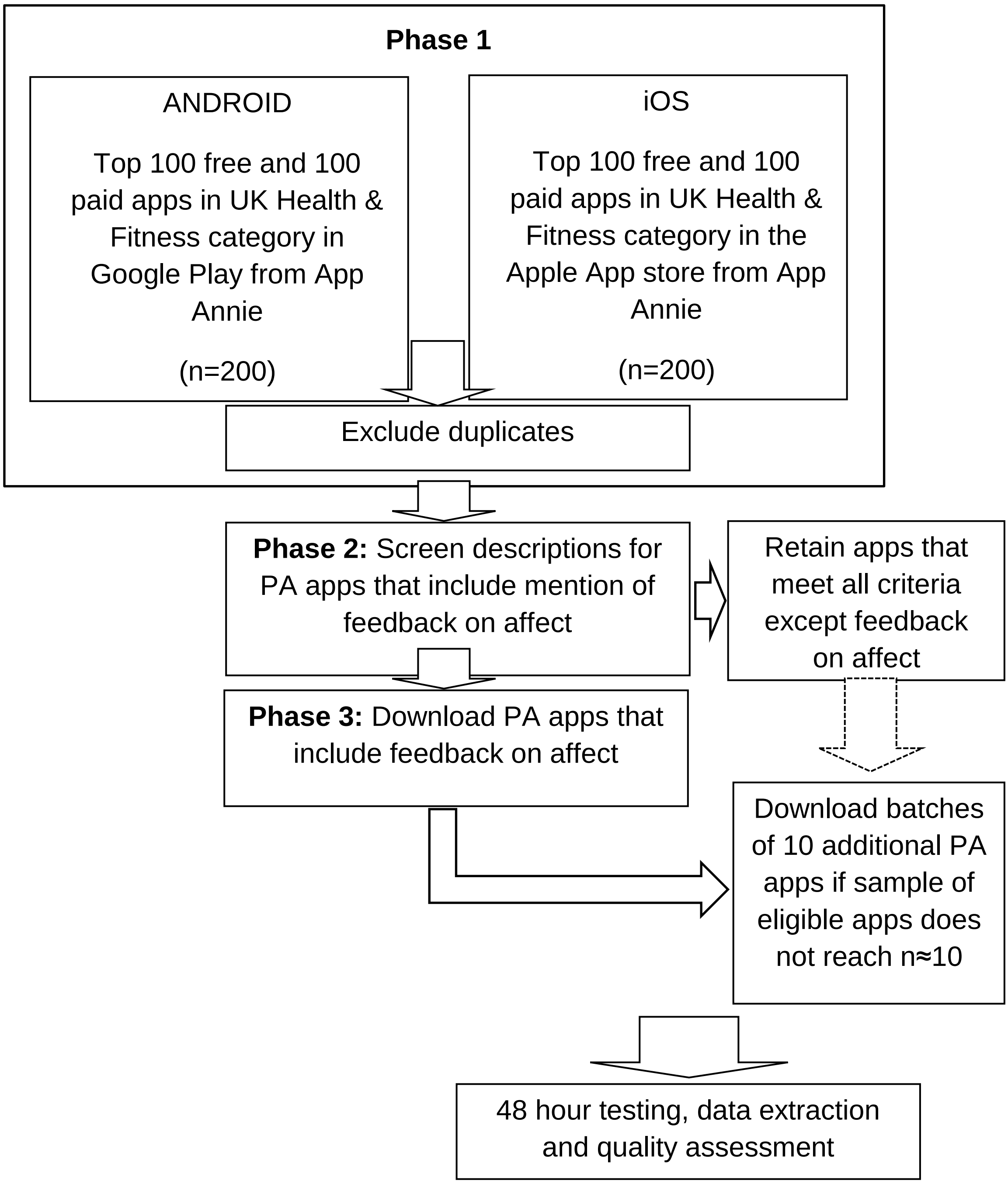
4.4.7.3 Phase 3 – Testing downloaded apps

Apps that were identified as potentially including feedback on affect were prioritised for download to confirm its presence. As communication with an author of a recent review suggested that few PA apps containing a mood component existed (Bondaronek, 2017; Bondaronek et al., 2018), a preferred sample of approximately 10 was the initial goal. When those apps prioritised for download did not allow that goal to be reached, the retained pool of PA apps from phase 2 were randomly sampled, in batches of 10, and downloaded until the sample size was reached and/or time to complete the study and financial resources were limited (see figure 9).

Apps were installed onto the same devices used for the systematic review. Apps were initially assessed by a single reviewer (LL), for approximately 1 hour, to determine presence of affect-based feedback, unless the app suggested that longer testing was required for identification. A second reviewer double-checked the apps based on the primary reviewer's decisions.

Apps that required additional linked devices e.g. wearables, were retained but were assessed based on app descriptions and in-app information only. Linked devices were not sourced due to financial constraints.

Figure 9 Flow diagram of identification of relevant apps



4.4.8 Data extraction

As reported in previous studies of public apps, following confirmation of feedback on affect (phase 3), app developers and websites were contacted/viewed for further evidence and information on their app (Knight et al., 2015). Using the same methods as study 1, databases were also searched for any relevant evidence base for the apps using the app name (see appendix 14 for strategies). Papers were not retained if apps were tested on ineligible populations as defined in sections 3.4.3, 4.4.2, combined with other intervention content without individual evaluation of the app, or were focused on the affiliated wearable without individual evaluation of the app. The apps and their additional evidence (websites and academic papers) were collated and together informed data extraction and quality assessment.

Apps were trialled for approximately two days in accordance with other app evaluations and recommendations, (e.g. Bardus et al., 2016; Bondaronek et al., 2018; Knight et al., 2015; Middelweerd et al., 2014) unless the app suggested that longer testing was required to identify feedback on affect and/or BCTs. The apps and any additional evidence were data extracted and quality assessed using a standardised, piloted, proforma consisting of the aforementioned tools and items (4.4.4). It was informed by the Template for Intervention Description and Replication (Hoffmann et al., 2014) and extracted data are summarised in table 13 (see appendix 15 for guidelines).

Table 13 Extracted data for public apps

Data type	Data extracted
Administrative details of app	Name, version, rating, developer, developer credentials, number of ratings, release date, last update, cost, platform, affiliations, technical aspects, whether extra devices were required/paired with app.
Targeted users	Diagnostic criteria, age, sex, socio-demographics, ethnicity.
App content and target	Size of app, security and privacy features, description, tailoring/personalisation, PA measurement, targeted activities/behaviours, inclusion of PA recommendations, type of affect and how it was captured, processed and fed back and how long it was available for, BCTs, BCTs for PA, use of other tools/materials to feedback affect.
App development	Involvement of users
Measures and results of quality assessments, most likely reported in extra evidence if available	Usage, if and how quality was assessed, results of quality assessments, whether or not quality issues were addressed
Quality results	As captured by MARS and TCS
Other	Miscellaneous comments from extra evidence
User reviews	Fifty most recent user reviews from the date the app was downloaded

BCT = Behaviour Change Technique, MARS = Mobile App Rating Scale, TCS = Theory Coding Scheme

During preparation for data extraction, it was discovered that two apps had automatically updated themselves and could not be downgraded back to the original version (Runtastic, Runtastic PRO). User reviews were subsequently re-extracted from the date of update based on the software version.

Piloting of the data extraction tool which included all of the quality assessment items, on one app by two researchers led to the following adaptations:

- Top 50 reviews were extracted separately by LL and not subjected to double extraction.

Two apps (One You Couch to 5k, Keep) had MARS and the BCTs independently extracted by two researchers. As the data extraction forms for the two studies were so similar, and no major concerns were raised for content beyond MARS and the BCTs, it was considered acceptable as well as pragmatic, for the rest of the extracted data for the two public apps to be double checked, rather than double extracted. One additional public app's data was double checked by an expert in health interventions (Runtastic Heart Rate PRO). These apps were chosen as they represented the range of apps: promoting PA only, promoting PA and other behaviours and a self-monitoring app alone.

4.4.9 Analysis and synthesis

Analysis and synthesis were identical to those performed for the systematic review apps (see section 3.4.11). The one exception being data from user reviews.

User reviews were thematically analysed. A combination of deductive themes and inductive themes were developed. Deductive themes were informed by a priori complaints, or overarching categories adapted for PA apps sourced from Khalid et al., (2014) and Milward et al., (2016) as well as derived from the 13 quality indicators defined in 3.2. Six phases of thematic analysis were conducted in accordance with guidance from Braun and Clarke, (2006), with themes discussed and reviewed for face validity within supervisory meetings. The variety and type of quality issues were of interest and only a limited number of app reviews were extracted, therefore it was not appropriate to perform content analysis or an assessment of inter-rater reliability. NVivo 11 was used to organise and code data. User reviews were also classified as either having a positive or negative sentiment, or both.

5. Combined results of studies

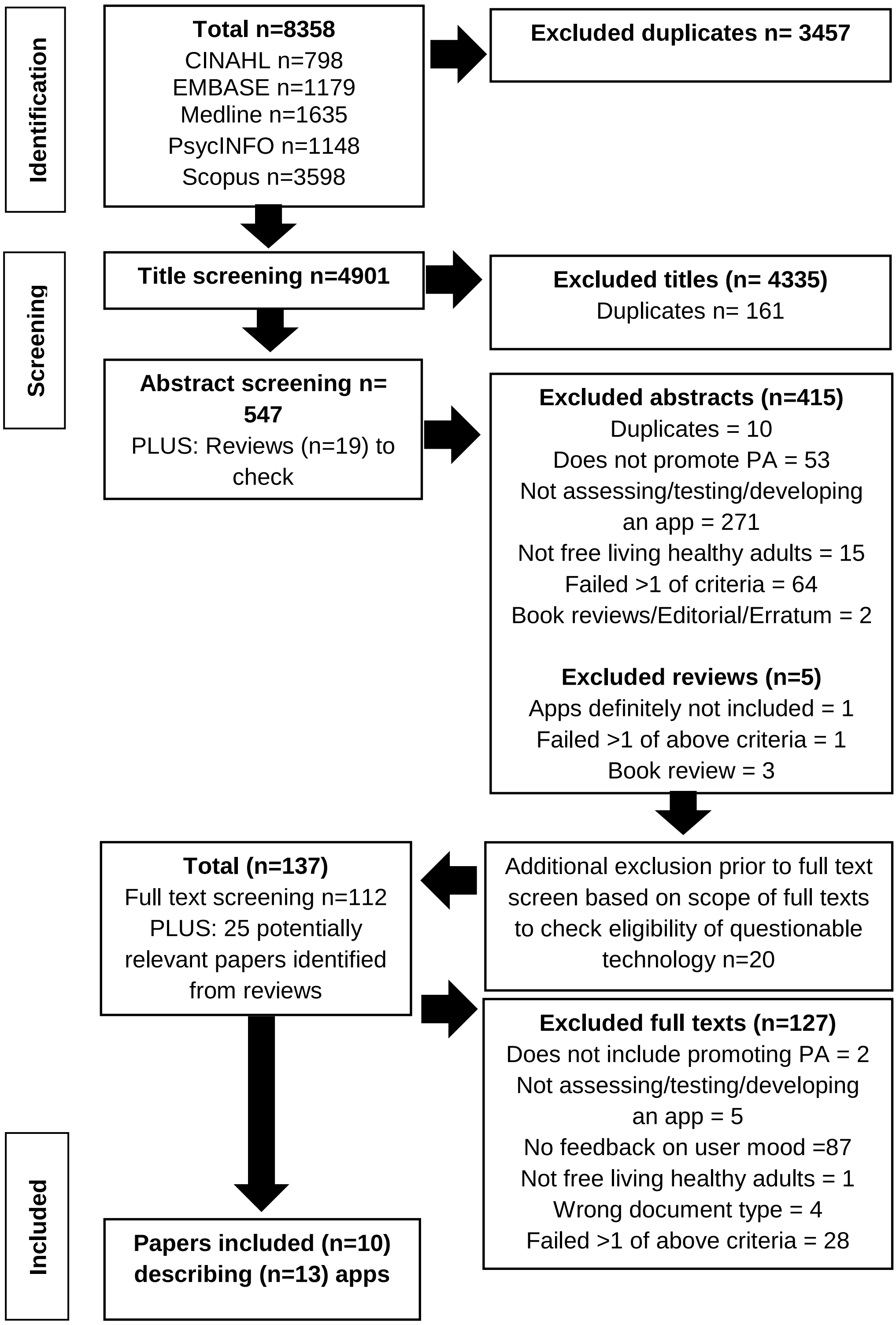
5.1 App selection

5.1.1 Apps identified by systematic review (study 1)

Initial screening identified 137 papers, including 25 potentially relevant additional papers found from relevant reviews that appeared in the original searches. Following full text screening, 10 papers describing 13 apps (Fanning et al., (2017) reported four apps) were found to be eligible for inclusion (Bentley et al., 2013; Brindal et al., 2016; Fahim et al., 2014; Fanning et al., 2017; Fernández et al., 2013; Forster et al., 2017; Harris et al., 2018; Hearn et al., 2014; Kim et al., 2015; Podina et al., 2017). Many ecological momentary assessment studies describing apps that captured PA and affect were identified. However, they were excluded as they looked for a relationship between the two variables, rather than aiming to promote PA (e.g. Lathia et al., 2017). One app was included despite it being a web app, rather than a native app, as the authors advocated its use on a smartphone (Fanning et al., 2017). Others had allowed access via any medium (computer, tablet, smartphone). One app was retained despite the age range starting in young adolescence as the users included those up to 29 years of age, despite using an iPod Touch as it used apps and was anecdotally considered a precursor to the iPhone (Kim et al., 2015). In addition it was one of few apps targeting a more at risk population One app that was initially included was subsequently excluded, as the only available information was from a short-form conference paper and no further details of the app were forthcoming from authors (Deline et al., 2012).

Four authors replied to requests for further information/evidence. Further details were provided for one app (Bentley et al., 2013). No additional papers were provided and no apps were made available or located in app stores. Eleven additional papers were found from searches, that provided extra evidence for the apps (Ali et al., 2015; Brindal et al., 2018; Cleland et al., 2013; Fatima et al., 2015; Idris et al., 2015; Lyons et al., 2014, 2017; Melton et al., 2016; Podina et al., 2018; Saleem et al., 2012; Tollmar et al., 2012). One additional paper describing one literature-based app (Up) and one public app (Fitbit) was identified during searches conducted for evidence for the public apps also (see 5.1.2 and figure 10).

Figure 10 PRISMA flow diagram



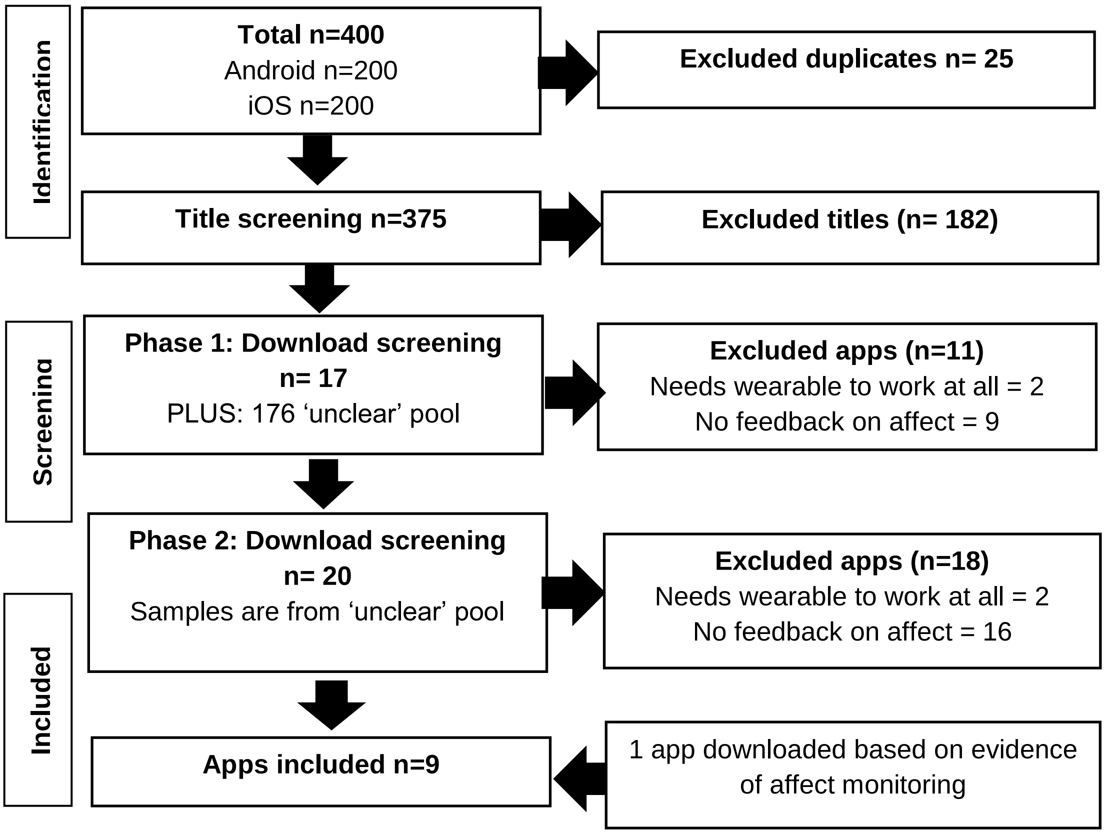
5.1.2 Apps identified by systematic evaluation of publicly available apps (study 2)

Four-hundred apps were extracted from App Annie on the 31st October 2018 (App Annie, 2010), plus an additional two paid apps, as it was discovered that two free apps appeared in the top paid Android list erroneously (FODMAP; Gluten Free Scan UK – Full – Coeliac healthy diet). These were taken sequentially from places 101 and 102 of the list. Following de-duplication, 375 apps had their titles and descriptions reviewed for eligibility, resulting in 17 apps for download and 176 apps where it was unclear if they incorporated affect. Descriptions of yoga apps were reviewed by LL twice - based on feedback from double screening - as they often referred to elements of wellbeing and mood. However, descriptions were sufficiently vague to suggest that these apps were ineligible. Downloading and assessing the apps resulted in 6 eligible apps – Runtastic, Runtastic Heart Rate PRO, Runtastic PRO, Period Tracker Clue, One You Couch to 5k and Samsung Health. One additional app from the ‘unclear’ pool, Runkeeper, was downloaded for assessment and subsequently included, based on evidence that it included an affective component (Baretta et al., 2018). An additional random sample of 10 apps from the ‘unclear’ pool was taken from each of the paid and free apps to try to bring the total for evaluation closer to 10. Two additional free apps were included: Fitbit and Keep, resulting in a total of nine apps, seven free (Fitbit, Keep, Runkeeper, Runtastic, Clue, One You Couch to 5k and Samsung Health), two paid (Runtastic Heart Rate PRO, Runtastic PRO).

Six app developers replied to requests for further information/evidence on eight of their apps (where one developer, Runtastic, created 3 eligible apps). While most referred to sections of the app website for further general information about the app, only one provided details of an academic study evaluating the apps quality (Clue, (Moglia et al., 2016)).

Twenty-six additional papers were found describing one app, Fitbit, (Altenhoff et al., 2015; Antón and Rodríguez, 2016; Bauer and Kriglstein, 2016; Beltrán-Carrillo et al., 2019; Bouts et al., n.d.; Crawford, 2013; Eisenhauer et al., 2017; Ellingson et al., 2019; Fallaize et al., 2019; Freis et al., 2018; Griffiths et al., 2018; Karduck and Chapman-Novakofski, 2018; Klenk et al., 2017; Klock and Gasparini, 2015; Lyons et al., 2014; Martinez-Nicolas et al., 2017; Miller et al., 2014; Papon et al., 2015; Pentakota et al., 2019; Pipitprapat et al., 2018; Poojary et al., 2018; Ramirez et al., 2016; Tong et al., 2018; Towler et al., 2018; Ullrich et al., 2016; Zhou et al., 2016). However, some were excluded due to reasons outlined in section 4.4.8. One of these papers also described one of the apps identified in the systematic review (Up). Four additional papers described five of the apps (Bondaronek et al., 2018; Direito et al., 2014; Middelweerd et al., 2014; Pagoto et al., 2013). Additional papers rarely reported the version of the app, therefore some evidence relates to different versions to those that were assessed for this thesis (see figure 11).

Figure 11 Flow diagram for public app identification



A combined total of 22 apps were included and evaluated. Apps from the first study are subsequently referred to as literature-based apps. Apps from the second study are referred to as public apps. The following sections will describe the results.

5.2 Study and app characteristics

5.2.1 Inter-rater reliability

Inter-rater reliability was assessed in a portion of apps (n=6, 27%). Statistics were calculated for Mobile App Rating Scale (MARS) scores and presence of Behaviour Change Techniques (BCTs), for each pair of independent double coders. The researcher who double checked the data found no MARS or BCT discrepancies, so analysis was only performed for the other coders.

5.2.1.1 Descriptive data

Negligible differences between coders were found for all other extracted data. Inter-rater reliability statistical analysis was not performed on this part of the data set.

5.2.1.2 Behaviour Change Techniques (BCTs)

There was moderate agreement on the presence of BCTs in apps identified from the literature between LL and coder 1, and LL and coder 2, resulting in the same statistic ($k = 0.789$, $p<0.005$, 98% agreement). Using the adjusted kappa, agreement was almost perfect between LL and both coders ($k=0.957$, $p<0.000$).

There was almost perfect agreement on the presence of BCTs in public apps between LL and coder 1 ($k=0.965$, $p<0.005$, 99% agreement) and weak agreement between LL and coder 2 ($K=0.526$, $p<0.005$, 88% agreement). Using the adjusted kappa, agreement was almost perfect between LL and coder 1 ($k=0.979$, $p<0.005$) and moderate between LL and coder 2 ($k=0.763$, $p<0.000$).

On further examination, it became clear that agreement was lower for coder 2 for the public app, as some app content had been missed by the coder. BCT agreement was considered acceptable for LL to continue single coding for the remaining apps.

5.2.1.3 Mobile App Rating Scale (MARS)

A weighted Cohens K was run to determine if there was agreement between the main coder (LL) and a second coder for each of 25 MARS scores. Statistics for all coder pairings indicated either weak or minimal agreement. Although actual agreement (where both coders provide the same numerical score) was poor, general agreement – or almost perfect agreement (whereby coders are within one score of each other) was observed to be high, as well as relevant given the linear Likert scales employed. Percentage general agreement has been used in place of inter-rater reliability statistics previously in health research, when kappa values were not representative, and for assessing app content (Hardeman et al., 2014; Middelweerd et al., 2014). Therefore, percentage general agreement was calculated (see table 14). General agreement was considered acceptable for LL to proceed with single coding.

Table 14 Inter-rater reliability percentage agreement for MARS

	Type			
	App 1 (systematic review)	**App 2 (systematic review)**	**App 1 (public)**	**App 2 (public)**
Coder pair	Pair 1	Pair 2	Pair 1	Pair 2
Actual % agreement	32	28	48	48
General % agreement	84	76	96	80

5.2.2 Descriptives

5.2.2.1 App study characteristics

The primary studies describing or evaluating apps identified in the literature were published between 2013 and 2018 and carried out across the developed world including Europe, USA, Korea and Australia. No apps were available to download and evaluate fully. Five apps had additional published evidence beyond author correspondence (Up, ATHENA, Health Mashups, Motimate, SIGMA). Study design and methods were mixed, with four studies reporting a randomised design. Sample size ranged from 7 to 196 depending on the type of design employed. Almost every app was quality assessed, but only half the apps subsequently addressed identified issues, although some authors proposed future changes. Some quality measures were reported as being assessed, but results weren't reported. This is due to using the definitions for the 13 quality criteria specified in Methods section 3.2.1, table 8, and the fact that some papers used different terminology or definitions of quality criteria between stating what was assessed and what the results of that assessment were. Only one app captured a facet of quality that was not included in the 13 criteria (Podina et al., 2017) (see table 15).

Table 15 Characteristics of studies identified through systematic review

App name and study citation (Name, author, year, title, country)	Availability of app and extra evidence	Study design & duration (type, design, duration)	Sample and number of groups	Quality assessment by authors (was it assessed /reported, were issues addressed)
Haptivity Forster et al., 2017. Using affective judgements to increase physical activity in British adults. Britain, UK.	App unavailable. Author correspondence.	Multi-method: qualitative and quantitative. Focus groups, co-design, process evaluation via email, discussions and questionnaires. Questionnaires, before and after pilot study lasting 4 months.	1 group provided participants for both qualitative and quantitative phases. Qualitative: n=16 initially, reduced to 9 and 7 respectively by final focus group. Quantitative: n= 7	Usage, functionality (outcomes not reported), acceptability, credibility, aesthetics (outcomes not reported) and effectiveness. Developed app iteratively based on feedback from users, but unclear what or how many changes were made.
MAPS (Multiphase Activity Promotion Study) Fanning et al., 2017. A smartphone "app"-delivered randomized factorial trial targeting physical activity in adults.	App unavailable. Author correspondence.	Quantitative, randomised four arm factorial trial lasting 12 weeks.	4 groups N=116 of which 13 lost to follow up.	Usage, acceptability, credibility, functionality, engagement. Ways to address issues were proposed but not enacted within paper.

Illinois, USA.				
Up app Harris et al., 2018. Enhancing psychosocial constructs associated with technology-based physical activity: a randomized trial among African American women. Georgia, USA.	App unavailable. 4 additional papers describing/evaluating app (Altenhoff et al., 2015; Lyons et al., 2014, 2017; Melton et al., 2016)	Quantitative, randomized two arm, pre-post follow-up design lasting 6 weeks with 2 months post-intervention follow up.	2 groups N= 71 of which 22 lost by 2 month follow up.	Usage, acceptability, functionality, credibility. No issues addressed.
Unnamed app Fernandez et al., 2013. Using smartphone bases biodevices for analysing physiological, psychological and behavioral user's habits. Spain.	App unavailable. No additional evidence.	Description of app and its design	NA	No
ATHENA apps Fahim et al., 2014. ATHENA: A personalized platform to promote an active lifestyle and wellbeing based on	One app available: Step Counter. 5 additional papers describing/evaluating app (Ali et al., 2015; Cleland et al., 2013; Fatima et al.,	Descriptive and qualitative. Describes components of app and a mock case study scenario.	NA	Acceptability (in app so outcomes not reported), credibility (of one component only). App adjusts recommended content

physical, mental and social health primitives. South Korea.	2015; Idris et al., 2015; Saleem et al., 2012)			for user based on user acceptability feedback
Health Mashups Bentley et al., 2013. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behaviour change. Chicago, USA and Stockholm, Sweden	App unavailable. Author correspondence and 1 additional paper describing/evaluating app (Tollmar et al., 2012)	Multi-methods pilot study consisting of qualitative design using interviews and grounded theory and quantitative field evaluation, lasting 2 months. Quasi-experimental trial study consisting of qualitative interviews, questionnaires and thematic analysis and quantitative analysis of usage and wellbeing outcomes lasting 90 days.	1 group for Pilot and Trial Pilot n=10 Trial n=60	Pilot - Usage, acceptability, credibility, engagement. Some issues identified across all quality themes were addressed before trial. Trial – Usage, credibility, acceptability and functionality. No further issues addressed following trial.
iN Touch Kim et al., 2015. Youth-centered design and usage results on the iN Touch mobile self-management program for overweight/obesity.	App unavailable. No extra evidence.	Mixed-methods. Participatory design (Design) and non-controlled feasibility (Implementation) study. Design consisted of discussions, focus	1 group for Design and Implementation studies Design n=10 Implementation n=34, with 10 lost to follow up.	Usage, acceptability, functionality, credibility, engagement, across both studies Issues identified during Design stage were addressed before

California, USA		group, user testing sessions, and interviews – full duration unreported. Implementation test consisted of individual case report within single group reporting usage and including interviews and questionnaires, during 6 months.		Implementation stage and proposed further changes after Implementation stage.
Motimate Brindal et al., 2016. Combining Persuasive technology with behavioural theory to support weight maintenance through a mobile phone app: protocol for the Motimate app. Adelaide, Australia	App unavailable. 1 additional abstract describing/evaluating app (Brindal et al., 2018)	Multi-methods, quantitative and qualitative. RCT assessing efficacy and feasibility, 24 weeks. User evaluations.	2 groups N=88	Usage and acceptability. No issues addressed.
Ngala Healthy You, Healthy Baby Hearn et al., 2014. Reaching perinatal women	App unavailable. No extra evidence.	Multi-methods. Qualitative development using focus groups, interviews and logico-	1 group N=196 consisting of 53 pregnant women, 67 postnatal women, 76 primary health care	Assessed usage. No issues addressed.

online: The Healthy You, Healthy Baby Website and App. Western Australia		inductive and gap analysis, and quantitative usage and fidelity assessment 1 year after launch.	practitioners.	
SIGMA - Self-help, Integrated, and Gamified Mobile-phone Application Podina et al., 2017. An evidence-based gamified mHealth intervention for overweight young adults with maladaptive eating habits: Study protocol for a randomized controlled trial. Romania	App unavailable. 1 additional paper describing/evaluating app (Podina et al., 2018).	Mixed-methods pilot study proposed over 5-7days. Quantitative RCT with placebo control proposed with 2 month duration and 3 month follow up.	2 groups proposed. Proposed N=104 to retain power with 40% attrition rate.	Usage, aesthetics, functionality, acceptability, engagement and credibility. Also assessed Player Experience. Proposed ways to address issues prior to efficacy study.
Descriptives	**Apps available: 0/10 (part of 1)** **Apps with additional evidence:7/10**	**Multi/mixed methods: 6/10** **Quantitative: 2/10** **Qualitative 0/10** **Other: 2/10**	**Group range:1-4** **Sample size range: 7-196**	**Apps quality assessed: 9/10** **Quality issues actively addressed/propose to be addressed: 5/10**

RCT = Randomised Control Trial

5.2.2.2 Public apps and additional evidence

Of the 9 public apps identified, all of them included an affiliated website, and all but one had additional evidence (see table 16). For the purposes of this thesis, additional evidence from now on will refer to academic papers, or published reviews of the pros and cons of the apps in general, as well as the app website.

Table 16 Presence of additional evidence and version number for public apps

App name (developer)	App version evaluated	Availability of additional evidence* and academic papers
One You Couch to 5k (Public Health England)	7.06	Yes, plus 1 additional paper evaluating app (Bondaronek et al., 2018)
Runkeeper – GPS Track Run Walk (Runkeeper)	9.8.3	Yes, plus 7 additional papers evaluating app (Crawford, 2013; Direito et al., 2014; Klock and Gasparini, 2015; Martinez-Nicolas et al., 2017; Middelweerd et al., 2014; Miller et al., 2014; Zhou et al., 2016)
Runtastic Heart Rate PRO (Runtastic)	2.6	Yes, plus 3 additional papers evaluating app (Bouts et al., n.d.; Papon et al., 2015; Pipitprapat et al., 2018)
Runtastic Running app & Mileage Tracker (Runtastic)	9.3	Yes, plus 8 additional papers evaluating app (Bauer and Kriglstein, 2016; Bondaronek et al., 2018; Klenk et al., 2017; Klock and Gasparini, 2015; Pentakota et al., 2019; Poojary et al., 2018; Towler et al., 2018; Ullrich et al., 2016)
Runtastic PRO Running Fitness (Runtastic)	9.3	Yes, plus 1 additional papers evaluating app (Antón and Rodríguez, 2016)
Fitbit (Fitbit Inc.)	2.92	Yes, plus 11 additional papers evaluating app (Altenhoff et al., 2015; Bondaronek et al., 2018; Eisenhauer et al., 2017; Ellingson et al., 2019; Griffiths et al., 2018; Karduck and Chapman-Novakofski, 2018; Lyons et al., 2014;

		Middelweerd et al., 2014; Pagoto et al., 2013; Ramirez et al., 2016; Tong et al., 2018)
Keep: Fitness & Workout Trainer (Keep Inc.)	1.28.1	Yes, no papers.
Period Tracker Clue: Period & Ovulation Tracker (BioWink GmbH)	5.7.0	Yes, plus 2 additional papers evaluating app (Freis et al., 2018; Moglia et al., 2016)
Samsung Health (Samsung)	6.1.1.001	Yes, plus 2 additional papers evaluating app (Beltrán-Carrillo et al., 2019; Fallaize et al., 2019)
Descriptives	**NA**	**Apps with academic evidence: 8/9**

* App website

5.2.2.3 General app characteristics

For the following analysis, the Up app was reported with the public apps, despite it being identified during the systematic review. This is because it was publicly available from the app store at the time the study was conducted, a version of the app is still present in the app stores, and it wasn't specially developed by the authors, although it has been discontinued and no longer functions correctly.

For ease of reporting, the following shortened versions of app names are used from now on: Ngala (Ngala Healthy You, Healthy baby), One You (One You Couch to 5k), Runtastic HR (Runtastic Heart Rate PRO), Keep (Keep: Fitness & Workout Trainer), and Clue (Period Tracker Clue).

Eleven apps were designed for the Android platform and four for iOS. Five were cross-platform and two didn't have a platform reported, both of which were found in the literature. The size and cost of public apps were readily available, unlike those for literature-based apps. Only two public apps required payment (Runtastic Heart Rate PRO, Runtastic PRO). Of the 22 evaluated apps, seven promoted PA only, while the rest targeted multiple behaviours, including one that was primarily focused on menstrual cycles and their symptoms and antecedents (Clue) (see appendix 16 for details of other behaviours targeted and types of PA targeted.) Apps varied in complexity and volume of components across both groups, with Haptivity, One You and Runtastic HR being quite simple and having little content, while MAPS, SIGMA, ATHENA, Samsung Health, Keep and Fitbit were more complex, featuring more content. All but one app tailored, or allowed users to tailor, content/settings based on user entered data or user preferences (Unnamed).

Five literature-based apps specifically targeted sedentary users (Haptivity, MAPS 1 - 4). Three apps were tested by or targeted users from Black, Asian and Minority Ethnic (BAME) groups (Health Mashups, iN Touch, Up) and four reported their participants as being predominantly white (MAPS 1 -4). Two apps targeted women (Ngala, Clue), but the latter deliberately used gender fluid language. Apps included a range of technical aspects defined by MARS, with four reporting a single aspect (Unnamed app, Motimate, Ngala, Up). Public apps tended to feature more technical aspects (see table 17 for details).

Table 17 Characteristics of included apps

App name, platform evaluated, size, cost	Targeted behaviour (PA only or PA+)	App description (what it does and how, tailoring or personalisation)*	Users (diagnostics, age, sex, SES, ethnicity, other)	Technical aspects (MARS)
Literature-based apps				
Haptivity iOS Size and cost not reported	PA	Reminds users to be physically active at a specific time and context each day, coupled with a photograph that the user has previously posted denoting the positive feeling they experienced during a previous session of PA. Each time they are active, users would be reminded by the app to take a photograph that denoted the positive feeling they were experiencing and post it to the app. The specific PA behaviour was intended to be defined by the user during set-up of the app (although could be modified at a later date). Users would also be able to view the photos at any time. This was intended to remind them of their previously positive experience of PA to serve as additional motivation to get active should they need it. The app was designed to be used in the short to medium term, until a habit for a specific, user- defined PA has been formed. When users first open the app they were asked which specific PA behaviour they want to make a habit, when they would like to perform the behaviour and what they were likely to be doing at this time (e.g. before lunch). Users were encouraged to post a photograph that captures the positive emotions they experienced during PA (with a brief comment about	Sedentary but wanting to change this. Owned compatible smartphone. 35-55 years old, mean age by second round of focus groups: 47. Men and women. 78% women by second round of focus groups. Third focus group, and testers, 5/7 women. All testers were employed. Recruited from community.	Has an app community. Sends reminders.

		why they felt good). Users were reminded to perform the behaviour at the time they had specified, and the reminder was accompanied by a previously posted photograph. Users' posts could be seen publicly and they could receive a restricted range of positive feed-back provided by other app users. Feedback options were restricted to ensure that detrimental or negative comments were not posted. Users able to personalise/choose from options.		
MAPS (1-4) Cross-platform web app Size and cost not reported	PA	All conditions (n=4) had four features. 1) Tracking activities, 2) Instant feedback, 3) bi-weekly feedback, 4) knowledge. 1) Tracking of activities: Users entered duration, intensity, enjoyment, extra notes. 2) Tracking integrated with instant feedback: Instant indication of progress within the week, and historically throughout the program. Graphical feedback. Diary provided detailed information of weekly and historical activities. 3) Biweekly feedback also integrated with tracking: Support emails with goal-setting reminders, activity summary, tailored. Weekly educational content, and progress toward weekly goal. Sunday text-messages served as a goal setting reminder. Thursday text-messages provided a motivational quote and summary information or a reminder to be active. 4) Knowledge: Weekly educational modules, including videos that unlock a quiz question. Answering the	Healthy, low active (self-report no more than 30 minutes MVPA on 2 or more days a week) but able to walk without assistance and PA not contraindicated. Owned compatible smartphone. 30-54 year's old, mean age 41.38 +/-7.57. 80% female (n=93). 77% married (n=89). 84% college educated (n=98), 52% earning at least $70,000 annually (n=60) at baseline. 87% white (n=101). Recruited from the community. (Harris et al., 2018)	Sends reminders (but via text). Needs web access to function (as it's a web app). Requires login.

		quiz question unlocked ''*support*'' content viewable at any time. **1) MAPS** • Recorded weekly goals in printed workbook. **2) MAPS + Goal setting** • Urged progression toward public health recommendations for PA. Distal and proximal, and aerobic, non-aerobic and strengthening goals were set. • Viewed information specific to progress toward weekly goals. • More rapid progression was recommended following successful weeks, guidance following unsuccessful weeks. Distal (maintenance) goals were set only once previous distal goals were achieved. • Reputable exercise resources to facilitate ongoing goal setting. • App guided users through entering goals into the module, but the first set of goals were set under the supervision of the research staff. • Displayed participant goal progress during the week, and provided motivational alerts when goals were successfully met. **3) MAPS + Point based feedback** • Recorded weekly goals in printed workbook.		

		• Delivered instant SCT feedback and incremental rewards, using a system of ''*program points''* (awarded for all in-app tasks), *''levels''* (awarded for accumulated points), and '*'badges''* (awarded for every two earned badges), with the intent of building and supporting self-efficacy. **4) MAPS + Goal setting & Point based feedback** • All of the above Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics and collected data.		
Unnamed app Android Size and cost not reported	PA+	Captures physical activity, mood, breathing/sleeping and socialising data via GPS, external breath sensor and related phone usage including calls, SMS and social media use. Sends data to server database for long term processing of adherence to health behaviours, with the eventual aim to get results that can improve user lifestyle. Provides feedback on behaviours through text-speak (physical activity – distance, time and speed, overlaid over a map) and mood (bar charts) and breathing (real-time trace graph). No tailoring/personalisation reported.	Young and adult people.	Needs web access for certain features.
ATHENA Android Not reported	PA+	ATHENA platform consists of multiple apps to identify underlying connections between physical, mental and social health primitives including: exercise routines, sleep, feelings and emotional states and outdoor visited places. Data are measured through pervasive	Any. Case study included 30 year old male, 156.5lbs, who wanted to adopt an active lifestyle and was	Allows sharing. Needs web access to function. Requires login. Sends reminders.

for entire platform, but step counter app: 2.33MB Step counter: Free		sensing devices and processed both online and offline by machine learning algorithms. On the basis of personal profile parameters, recommendations and interventions are provided to the users. Feedback from users alters recommendations for wellness services. On the basis of the user's lifestyle, platform makes exercise, food and social networking recommendations. ATHENA can provide recommendations, such as "*do not do physical exercise or take coffee X hours before sleeping*". In this way, a user can change his lifestyle while adopting an active lifestyle. ATHENA provides web interfaces that give access to visualizing the subject's daily, monthly and weekly physical routines through our graphical widgets. ATHENA platform will compare daily routines to another group that is comparable and report this to the user. Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics, preferences and collected data.	an active social media user recruited from the community.	
Health Mashups Android Size and cost not reported	PA+	Contextual and wellbeing data from multiple data sources collected, ideally daily, (automatically sensed location, weather, and calendar free/busy hours via phone; sensors indicating step count and sleep (Fitbit) and weight (Withings scale)). Users can manually log food, exercise, mood (final version only) and pain. Web server conducts statistical analysis and feeds back significant user-specific observations of relationships between and within sensor data points, e.g. links	Adults (22-65 Pilot, 22-65 Trial) ranging in weight between underweight to overweight (150lbs). Men and women of diverse occupations and education levels.	Sends reminders. Needs web access for certain features.

		between environmental context and behaviours. Both deviations and correlations are calculated and fed back, daily, by day of week, weekly and monthly. Feedback delivered to app daily including plain-text confidence of the relationship and an in-app link to a mobile-website showing graphs and data for user exploration of data. Feedback is a deliberate mix of items, correlations and deviations to ensure daily changes over time. Graphs showed all data points for observation per day or plots of sensor vs. time. Goal is to encourage reflection, remind users of what affects their wellbeing and therefore encourage behaviour change. Status bar notification added for when new statistical observations were available. Users able to personalise/choose from options. App-based tailoring and recommendations based on collected data.	Pilot included participants from diverse cultural and family backgrounds and situations, with limited technical literacy. Recruited from community and extended networks of research team. Trial included participants living in range of accommodation from subsidised housing to large suburban homes. Trial included participants with varying technical literacy. Recruited from community.	
iN Touch iOS Size and cost not reported	PA+	A mobile application with pre-existing trackers incorporated modified food and exercise trackers and additional mood and socialising trackers as well as modifications to app navigation. New features allowed a health coach to manage users and included development of a web site. The new features included (1) compilation of a custom configuration of trackers	Overweight/obese with a BMI more than or equal to 85th percentile for gender and age. No significant medical conditions. 13-29 years old across	Allows password protection. Needs web access to function. Sends reminders (but not clear if app-based).

		and features to be distributed to all study participants, (2) a dashboard for viewing the status of participants and drilling down to individual level ODL data and (3) uploading of batch ODL reports to an electronic health record. Following motivational interviewing with a health coach, users developed his/her own health goals and action plans for accomplishing the goals. The app tracked ODLs, and allowed review of recorded data as well as access to photos and notes that were inserted into a daily journal along with tracker entries. The entries from each tracker were inserted into the daily journal at the time they were entered and participants could add any text notes or pictures. Users could add additional behavioural trackers to the app. The full view of the daily journal was accessible on the web but not in the mobile application. The health coach and clinicians could review ODL data, meet with users and/or communicate concerns based on data in person. Users able to personalise/choose from options.	Design (15-29 years old) and Implementation (13-24 years old). Mean age 18 years during implementation. 6 women in Design group, Implementation comprised 73.5% women. Low income, urban. Several had one or more children. Design - 6 race/ethnic categories. Implementation – Sought minorities. 16 (47.1%) Hispanic/Latino, 7 (20.6%) African American, 5 (14.71%) Asian/Pacific Islander, 2 (5.8%) mixed/other, 4 (11.7%) unknown. Range of experience using smartphones from never to all the time. One participant did not use internet for at least once a week. Recruited from community,	

			University and school and hospital clinics.	
Motimate No details reported	PA+	Aims to improve a user's personal coping resources through basic behavioural therapy techniques that encourage workshopping of resources to deal with different moods and stresses, while also providing weight, diet and exercise monitoring and utilising immediate two-way feedback. Includes prompts to enter and review data. Feedback on weight and weight gain including tailored motivational messages depending on current weight as well as coloured weight depiction. Wellbeing monitored so users can identify links between emotional states and behaviours. Diet and exercise monitoring including personalised daily targets and recognition when daily allowance sustained/exceeded. Automatic text feedback on exercise records. Summary graphs provided to reflect on and identify potential patterns. Behavioural reviews provided weekly, including feedback on good and poor areas. Action planning used for diet monitoring, when consistently over/under recommendations, predefined actions are listed to facilitate 'getting back on track' and meeting targets. Coping tool includes psycho-education and guidance on maintaining positive wellbeing - only appears when large positive or negative change in mood or stress detected. Then *'hassle interface'* and coping strategies are presented. If a positive change in mood is identified, users will be asked what strategy they used to change it. Feedback	Adults who had lost and maintained at least 5% of their body weight within last 2 years. 75% (n=66) female at baseline, 69.3% (n=61) at end. Recruited from community including health clinic.	Sends reminders.

		on mood changes and strategies used will be provided on home screen. Daily motivating messages transition from general motivation to action planning to coping planning. Personalised feedback provided in monitoring tools. Administrator gets email notification of need for further intervention. App-based tailoring and recommendations based on user characteristics and collected data.		
Ngala No details reported	PA+	Self-assessment tool to track maternal lifestyle behaviours and weight during pregnancy and the first 18 months of motherhood. Generates supportive tailored feedback and tips on how to make improvements. Information on height, initial weight, stage of pregnancy or postpartum collected to allow appropriate feedback and tailored information to be provided. Provided perinatal mothers with a personalised, interactive tailored resource with parent focused, brief advice relevant to their stage of pregnancy and lifestyle assessment, goal setting, and monitoring. The content was clinically endorsed and available via app and more detailed factual, practical, and localised information on a website. App-based tailoring and recommendations based on user characteristics and collected data.	Development - pregnant, post-natal women, PHCPs. Final app/web - perinatal/pregnant women and those within the first 18 months of motherhood. Recruited from hospitals and community	Requires login.
SIGMA Android Size and	PA+	The SIGMA app was designed as a serious game and intended to work as a standalone app for weight maintenance, or alongside a calorie-restrictive diet for weight loss. It uses a complex and novel scoring system that allows points earned within the game to be	Overweight (25</= BMI </= 29.9). Non-clinical maladaptive eating habits embodied by food cravings, binge eating,	Allows sharing. Has an app community. Needs web access to function.

cost not reported		supplemented by points earned during outdoor activities with the help of an embedded pedometer. The SIGMA intervention is designed to accommodate four mHealth modules. Additional information is also accessible via the study's website. The app offers the option of social media sharing as well. 1) Psycho-education module: includes information about app, behaviours, behavioural and cognitive styles, coping strategies. Daily tips and messages delivered. 2) The gamified intervention module: incorporates two sub - modules: the explicit cognitive-behavioural intervention (SIGMAe) and the implicit attention-training intervention (SIGMAi). SIGMAe: users must help characters choose healthy options. Points earned. SIGMAi: memory-based game to *"remember"* healthy options. Points earned. 3) The Crisis and Relapse prevention module: motivational messages or cognitive-behavioural coping strategies, dependent on the type of encountered issues. Provides relaxing breathing exercises. Can act as distraction. 4) The self-monitoring, feedback, and evolution module: Baseline tailoring to identify vulnerabilities. Users self-monitor their own	emotional eating and maladaptive behaviour and cognition towards food. No medical conditions that contra indicate physical activity and dietary recommendations, eating disorders, depression or psychotic disorders, appetite suppressing medication or enrolment in other weight management programmes. Android-compatible smartphone that is able to connect to the Internet. 18-35 years old. Men and women Proposed recruitment from university and online communities.	Allows password protection.

		eating and physical activity patterns. App offers personalised tips and feedback. Monitoring of physical activity is aided by an embedded pedometer, comparing the user's daily performance with a daily suggested target, providing feedback. App also monitors how well the participants apply the CBT principles to real-life situations via ABC diary that focuses on understanding the Antecedents and Consequences of maladaptive Beliefs. The diary provides healthy alternative ways of thinking and coping tips, or allows the user to write some personally motivational healthy statements. Intuitive feedback, in the form of charts, regarding the cognitive, behavioural, and emotional indexes of progress are provided. Points system: Three sources of earning points: (1) SIGMAe, (b) SIGMAi, and (c) the pedometer. Earned points help the user reach a higher mastery level. A calorie counter is also available, but no points are earned for its usage. Users able to personalise/choose from options – minimal. App-based tailoring and recommendations based on user characteristics and collected data.		
Descriptives	**PA focus: 5/12** **PA+: 7/12**			**Allows sharing: 2/12** **Has an app community: 2/12**

				Needs web access to function:9/12 **Allows password protection: 2/12** **Requires login: 6/12** **Sends reminders: 7-9/12$^{\$}$**
Public apps				
One You Android 104MB Free	PA	A user chosen motivational trainer guides the user through 27 runs over 9 weeks of increasing intensity to reach a goal of running 5k, for 30 minutes, 5 times a week. Reminders for runs can be set by the user. Each run programme is described in text and verbally. User is asked how they feel before and after run with emoji's. Users able to personalise/choose from options.	Total beginners or anyone without conditions that contraindicate running.	Has an app community. Sends reminders. Needs web access for certain features.
Runkeeper Android 76.09MB Free	PA	Can track real-time activity, or can manually add activity and set own activity goals/targets. Receive badges and summaries for activities over time. Includes training plans e.g. My First 5k, with new sessions that are unlocked as others are completed and include an audio coach. Gives feedback on activities as well as goals. User feed allows summaries of activities to be shared with friends if desired and potentially receive encouragement. Can also see friend's activities there. Can join single or group challenges. Can receive notifications.	Those wishing to improve fitness, across all skill levels. At least 13 years old+	Allows sharing. Has an app community. Allows password protection. Requires login (but stays logged in). Send reminders. Needs web access for certain features.

		Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics.		
Runtastic HR iOS 53MB £1.99	PA+	Measures heart rate with camera using finger, then using icons, asks what user was doing at time of measurement: general, resting, before or after sport or at max HR. Also asks *"how do you feel?"* captured by one of five emoji's. Shows heart rate on a coloured slider. Can add notes and share on social media. Can view historical heart rate measurements and filter for certain types e.g. only resting HR measures. Can select previous measures for full details of mood, measurement type, and notes. Includes notifications, generally about social interactions, goal and challenge progress, syncing of wearables and event reminders. Can't edit entries. Users able to personalise/choose from options. App-based tailoring and predictions based on user characteristics and collected HR data.	Any skill or technology-literate level. At least 16 years old+	Allows sharing. Allows password protection. Requires login (but can stay logged in). Sends reminders.
Runtastic Android 83.13MB Free	PA+	Records running activities and calculates distance, calories and average pace as well as mapping the run. Can include music and voice coach. Following completion of a run, user can add a photo, rate how it made them feel using an emoji, describe the terrain using a picture emoji, list the type of shoes user ran in and add notes and heart rate. Automatically captures the weather. Can add range of different activities, select workouts in terms of goals for distance, duration, and customise displayed values during activity. Can	Any skill or technology-literate level.	Allows sharing. Has an app community. Allows password protection. Requires login (but can stay logged in). Sends reminders.

		share activities. Provides feedback on runs and all activities completed in summary form and can view individual sessions for full information. Can see weekly/monthly/yearly/all statistics. Can set goals and receive updates on progress, view presence on leader boards curated by user by inviting friends. Can sign up for challenges, view news from Runtastic, and join communities. Users able to personalise/choose from options.		
Runtastic PRO Android 83.23MB £4.99	PA+	Same content as Runtastic Running app & Mileage Tracker, but includes ability to auto pause runs (i.e. activities automatically pause when user does) and user can choose routes for activities. User can also select more workouts including challenges against previous performance, interval training and target paces. Additional music functionality. Can specify distance over a specific duration and calories burned as part of workout goal. User also receives data on maximum elevation. Users able to personalise/choose from options.	Any skill or technology-literate level.	Allows sharing. Has an app community. Allows password protection. Requires login (but can stay logged in). Sends reminders.
Fitbit Android 104MB Free	PA+	Tracks steps, exercise, sleep, fertility/period, weight, water and calories as desired and gives feedback on them all both daily and over time. Allows challenges to be completed individually or in groups/competitions. User can collect treasures and badges. Includes a community where user can share progress and receive cheers, join groups, add friends. Can set goals. Fitbit coach app is signposted and can be downloaded for free and gives access to training plans.	Any. At least 13 years old+	Allows sharing. Has app community. Allows password protection. Requires login (but stays logged in). Sends reminders. Needs web access

		Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics. App-based predictions based on collected data.		for certain features.
Keep iOS 88.5MB Free	PA+	Tailored to experience level and goals. Users receive coins for completed activities. App recommends training plans. Customisable plans available, goals can be set. Programs are unlocked as coins are earned but free plans are available. Social media sharing and cheers can be received. Information provided about exercise such as exercise preparation and potentially experienced feelings. Rewarded with trophies at the end of training sessions. Asks how user feels after workout. Displays feedback and details of workout session. Gives training feedback over time, and provides comparisons to other users. Allows calorie logging and provides workout/diet recommendations. Can manually add or track workouts. Articles and training videos available. Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics.	Any skill level. At least 13 years old+	Allows sharing. Has an app community. Allows password protection. Requires login (but can stay logged in). Sends reminders. Needs web access for certain features.
Clue Android 39.30MB Free	PA+	Allows data to be logged manually and daily for flow, pain, emotions, sleep, sex, energy, exercise, as well as numerous other period, body, vitality, activities and medical characteristics. Builds up pattern of characteristics and patterns of your cycle to predict next period or fertility and allows analysis of characteristics collected. Can add reminders for when	Women, teens and people. At least 13 years old+, pre-menarche to menopause.	Has an app community. Allows password protection. Sends reminders.

		next cycle is likely to begin, whether it's late, fertility and ovulation, warnings for PMS and breast self-exams, temperature data to predict fertility. Users able to personalise/choose from options. App-based predictions based on collected data.		
Samsung Health Android 141MB Free	PA+	Tracks steps, active time, exercise, heart rate, stress, water intake, oxygen saturation, blood pressure, caffeine intake, sleep, blood glucose, weight and weight management practices, nutrition and ongoing exercise programmes. Gives feedback on all and can set targets/goals. Can provide insights across these tracked activities. Can tailor to user based on demographics including activity level. Can earn rewards (badges), achieve personal bests and receive weekly summaries. Can compare behaviour to other users or friends, join challenges, contact GP, and discover health related articles, programmes and products. Articles are not created by Samsung. Users able to personalise/choose from options. App-based tailoring and recommendations based on user characteristics.	Any. HR ranges from <1 to >80 years old.	Allows sharing. Has an app community. Allows password protection. Requires login. Sends reminders. Needs web access for certain features.
Up iOS and Android Not reported Not explicitly	PA+	The UpBand is a commercial accelerometer that links with the Up app to give the participant feedback about physical activity (total number of steps, active time, longest active time, longest idle time, intensity) and sleep. The UpBand also gives feedback on mood and nutrition. When users reach their daily step goal the Up app congratulates them. Finally, if they have accomplished daily or weekly exercise goals, user's	New generation smartphone owners with an email address. 18-24 years old, ages ranged from 18 to 23 years (M = 19.83 years, SD = 1.64). Women only.	Has an app community.

reported, but free version available		receive additional recognition by the Up app. (App may also incorporate a SmartCoach which adjusts content based on user behaviour, but details were unclear.) Social support from friends using the UpBand and Up app allows users to provide and receive feedback from their peers who also utilised the UpBand and Up app. Users able to personalise/choose from options. App-based tailoring and recommendations based on collected data and comparisons to general population.	College students, recruited from a University, freshmen (n = 18), sophomores (n= 13), juniors (n = 5), seniors (n = 14), and graduate students (n = 5) Varying weights and BMIs. African American.	
Descriptives:	**PA focus: 2/10** **PA+: 8/10**			**Allows sharing:7/10** **Has an app community: 9/10** **Needs web access to function: 5/10** **Allows password protection: 8/10** **Requires login:7/10** **Sends reminders:9/10**
TOTAL	**PA focus: 7/22** **PA+: 15/22**			**Allows sharing:9/22** **Has an app community: 11/22** **Needs web**

				access to function: 14/22 Allows password protection: 10/22 Requires login:13/22 Sends reminders:16-19/22[$]

*Where possible, text is verbatim or reduced for length verbatim, from papers.

[$] In some instances it was unclear whether the app provided the reminders, in which case a potential minimum and maximum total is provided.

ABC = Antecedents, Beliefs and Consequence's diary, BMI = Body Mass Index, CBT = Cognitive Behavioural Therapy GP = General Practitioner, GPS = Global Positioning System, HR = Heart Rate, M = Mean, MVPA = Moderate to vigorous physical activity, ODL = Observations of Daily Living, PA= physical activity targeted by app, PA+ = physical activity and other behaviours targeted by app, PHCPs = Primary Health Care Providers, PMS = Pre-Menstrual Stress, SCT = Social Cognitive Theory, SD = Standard Deviation, SMS = Short Messaging Service.

5.3 Characteristics of feedback on affect

Six apps gave feedback on PA-contingent affect, affect specifically related to physical activity. Two apps captured affect as part of a female health component (Fitbit, Clue). One used objective measures to record affect (ATHENA) and one used photos to depict affect (Haptivity). The rest used sliding scales, emoji's or icons, questions, or lists of emotions with two apps using validated self-report measures of affect (Unnamed, Health Mashups). Most apps reported affect in the context of PA, either a specific activity session, or in the context of PA as well as other behaviours in general. Just over half (n=13) performed additional processing on affect data to make it more salient or meaningful. The exceptions were literature-based apps and ranged from including a reminder to be active with the affect feedback (Haptivity), to reports of statistical correlations between affect and behaviours, showing patterns of behaviour and affect (Health Mashups). Where reported, feedback on affect appeared to be available for the duration of use of the app. Many apps provided feedback on affect immediately after data entry by the user or on a daily basis (n=12) (see table 18, and appendix 17 for screenshots).

Table 18 Characteristics of affect capture and feedback provided by the apps

App name (downloaded or not, target behaviour)	Type of affect captured/fed back	Method of affect capture	Method of affect feedback	Duration of affect feedback availability
Literature-based apps				
Haptivity (Not available for download) PA	PA-contingent affect.	Self-monitored via users posting a photograph that captures the positive emotions they experienced during PA and a brief comment about why they felt that way.	Reminder to perform the behaviour was accompanied by a previously posted photograph.	Duration of app use.
MAPS (Not available for download) PA 1 2 3 4	PA-contingent enjoyment.	Self-monitored via sliding scale of enjoyment experienced from PA session. Captured in the same screen as activity details are entered. Limited details about scale: mid-point is *'neutral',* high point is *'really enjoyed it'*.	Instantly fed-back with PA session data. Available in weekly average in the same screen as weekly minutes of PA, intensity average and activities. Only includes written text e.g. *"Your average enjoyment; 4.00"* with a smile emoji next to it. Available via automated weekly email feedback e.g. *"Your average enjoyment was: 4.25 (somewhat enjoyed it)"*	Not reported.
Unnamed app (Not available for download)	General affective state including components	Self-monitored via Profile of Mood States (POMS) consisting of 15 questions on 0-10 scales where 0 means not at all and 10 very. 3 questions per component. Questions	Processing unclear, but states that: app will forward physiological, psychological and behavioural data obtained to a remote server for long term processing. Feedback via app	Not reported.

PA+	of stress, depression, hostility, vigour and fatigue.	are launched by the app at the beginning and end of respiratory acquisition.	includes bar chart of 5 components of mood and a result graph of a completed mood test. Feedback provided within an app that also collects PA data, but doesn't appear to be combined with PA data at this point, but may be in future.	
ATHENA apps (Not available for download) PA+	Emotions - happy, sad, angry, neutral.	Audio-based emotion recognition using the embedded audio sensor of a smartphone – Play Emotion app.	Bar graph for each mood and its overall duration. Mental health tab in personal profile where data is stored and shown. Mood is summarised in daily routine page, in the recommendations tab of the website per day by neutral, happy and angry. Can select weekly, monthly and yearly summaries. Mood comparisons to comparable groups using bar graph of duration of angry, happy, neutral, sad.	Duration of app use.
Health Mashups (Not available for download) PA+	Trial only. Generic daily affect.	Self-monitored via 7 point sliders corresponding to 4 standard measures of mood: happy/sad, tired/awake, unwell/well, and tense/relaxed.	Along with rest of data, it was statistically analysed at the end of each day and only statistically significant findings (deviations or correlations) within or between mood data were fed back to the user via the app, mobile website and computer website (unclear if both websites available for full trial). Daily, day of the week, weekly and monthly analysis performed. Feedback included examples such as: *"You are*	Duration of app use.

			happier when you walk more" and *"On weeks when you are happier you walk more (quite likely)"* provided in text form with a natural language confidence indicator based on a p value. This was available in the app and by clicking on the sensor for mood or PA, all the (statistically significant) observations linked to that sensor were available. A bigger graph of the sensors recent values could be seen in the app or mobile website. The computer website also provided this information at the end of the trial. Mood data were captured along with PA data and feedback on both were available in the app.	
iN Touch (Not available for download) PA+	Emotions - anger, happiness, sadness, stress.	Self-reported via sliding scale from None to Extremely. Unclear if all need completing or not.	Available in app and website - website shows full daily journal and individual trackers or days with slider visible and context or pictures if added, when it was reported and recent entries that may have included a mood measure. Less detailed data reported in app.	Duration of app/website use.
Motimate (Not available for download) PA+	Generic and momentary feelings and stress.	Self-reported via a single item: *"How are you feeling at the moment?"* Captured pleasant and unpleasant moods and those with high or low arousal. Emotional intensity from 1-	Mood will be shown on summary screen, including strategy that resulted in a positive change in mood (coping strategy) or plan to target hassle that induced negative mood for next time.	At least at daily and weekly intervals for the current day/week.

		10 as well as the location and time of data capture. Included reporting details of the hassle that induced the negative mood change and users' selected coping strategy for next time, or strategy that induced positive mood change. Stress measured through intensity bar.	Includes time and location of mood capture. Feedback via weekly review also showing number of good mood days, when user was at happiest, number of large mood changes and an encouraging tip on coping strategies for mood. The weekly feedback is reported along with the weekly feedback for diet and exercise and includes an area to focus on.	
Ngala (Not available for download) PA+	Emotions.	Self-assessment tool – no further details provided.	Generates supportive tailored feedback and tips on how to make improvements. Within the context of an app that also promotes PA.	Not reported.
SIGMA (Not available for download) PA+	Emotional reactions after eating.	Self-reported via ABC diary capturing emotions as well as cravings, sabotaging thoughts that preceded and ensued from their eating behaviour. Users can choose from a list of emotions, rate their intensity, or fill in some of their own.	An automatically generated graph indicates whether a change in time occurred in any of the emotional variables and pinpoints triggering/ problematic situations. SIGMA provides chart-based feedback, regarding emotional indexes of progress as compared to the user's baseline level.	Not reported.
Descriptives:	**PA-contingent: 5/12 Generic:6/12**	**Self-report: 11/12 Objective:1/12**	**Processed feedback beyond just entered mood data: 10/12* As entered: 1/12 Other/Unclear: 1/12**	**Range: Daily – Duration of use where reported**

	Other:1/12			
Public apps				
One You (Downloaded) PA	Both general affect and PA-contingent affect.	Using emoji's before (*"How do you feel?")* and after a run (*"How do you feel about today's run now?")*	Post-run mood emoji displayed in summary description of run when viewed later.	Post run mood - Duration of app usage unless app reset. Pre-run mood never available.
Runkeeper (Downloaded) PA	PA-contingent.	Once activity completed, user can respond to *"how did this run feel?"* with emoji's. Also asks verbally how user feels during My First 5k training coach session.	Can view emoji and an additional descriptor for emoji (e.g. ok, great), in summary of completed sessions on app only.	Duration of app use.
Runtastic HR (Downloaded) PA+	Affect at the time of heart rate measurement.	After recording HR, user can respond to *"how do you feel?"* with one of five emoji's, big smile, small smile, straight line mouth, small frown, big frown with plaster over one eye. No descriptors. In website, however, emoji's described as awesome, good, so-so, sluggish, injured respectively.	Able to see emoji's in individual HR recording session history under *'Feeling'* with the chosen emoji in colour compared to the others in grey.	Duration of app use.
Runtastic (Downloaded) PA+	PA-contingent affect.	Once activity completed, user can select one of five emoji's, big smile, small smile, straight line mouth, small frown, big frown with plaster over one eye depicting awesome, good, so-so, sluggish, injured respectively. When editing however, mood is described	Can view emoji and descriptor for emoji in summary of completed sessions.	Duration of app use.
Runtastic PRO	PA-contingent			

(Downloaded) PA+	affect.	as *'Feeling'*.		
Fitbit (Downloaded) PA+	Period related affect.	Female Health tracker allows users to log ‘happy’, ‘energised’, ‘fatigued’ or ‘anxious’ using emoji’s on daily basis.	Selecting the day of data entry on the calendar, within the Female Health tracker component of PA app, shows mood data. Can view bar chart of moods logged over the cycle under *‘trends’* option.	Duration of app use.
Keep: (Downloaded) PA+	PA-contingent affect.	After an activity user responds to *“how do you feel?”* using a slider from hard to light - too hard for me, tough workout, can barely talk; feeling just right, easy for me, too easy.	Completed activity records and user *'training feedback'* i.e. the mood descriptor.	Duration of app use.
Clue (Downloaded) PA+	Generic emotions and *“Mental”*.	Daily, users can choose between ‘happy’, ‘sensitive’, ‘sad’ or ‘PMS’ (pre-menstrual stress) icons to log emotions. For *‘mental’*, users can pick between ‘focused’, ‘distracted’, ‘calm’ and ‘stressed’ icons.	Can view daily data via the calendar, with chosen icons highlighted. Full calendar view shows day’s data collected via an orange square. This data is alongside green squares for exercise, as well as other recorded data. Under analysis tab, users can view when they felt different emotions or mental categories and can see the days they were reported, plotted along a line depicting the users latest menstrual cycle. Not shown with PA, but PA can also be viewed the same way.	Duration of app use.

Samsung Health (Downloaded) PA+	Generic mood. Stress.	Mood captured via emoji's after HR, stress and oxygen saturation measurements. Users enter what they were just doing when HR was taken: activities or mood (in love, tired, unwell, excited, surprised, and sad). Stress captured objectively via HR monitor using phone's camera. In same measurement page users can also enter current status (neutral, happy, sad, in love, tired, sick angry, fearful, excited, surprised emoji's).	Reported back with HR/Stress/Oxygen saturation measures. Shows the emoji without its descriptor, for that measurement.	Duration of app use.
Up (Available version not functioning) PA+	Not reported.	UpBand app. Self-report response to *"How do you feel?"* Method of data entry not clear.	The UpBand gives feedback on mood. No further details reported.	Not reported.
Descriptives:	**PA-contingent: 4/10** **Generic: 2/10** **Both:1/10** **Other/Not reported: 3/10**	**Self-report: 9/10** **Objective: 0** **Both:1/10**	**Processed feedback beyond just entered affect data (e.g. as graph, with reminder): 3/10*** **As entered:6/10** **Other/Unclear: 1/10**	**Range: Duration of use, where reported**
TOTAL	**PA-contingent:**	**Self-report: 20/22** **Objective: 1/22**	**Processed feedback beyond just entered affect data (e.g. as graph,**	**Range: Daily – Duration of use**

	9/22 Generic:8/19 Both: 1/22 Other/Not reported: 4	Both:1/22	with reminder): 13/22* As entered: 7/22 Other/Unclear: 2/22	where reported

* Includes those apps that combine affect feedback with a reminder or convert affect data into a graph, for example

ABC = Antecedents, Beliefs and Consequence's diary, HR = Heart Rate, PA= physical activity targeted by app, PA+ = physical activity and other behaviours targeted by app, PMS = Pre-Menstrual Stress

5.4 Quality of apps

5.4.1 Acceptability/user perceptions

Quality indicator	Summary definition
Acceptability or user perceptions	Positive and negative feedback or recommendations from users on the app content e.g. preferences or recommendations for information, tone or features. Can include user ratings or reviews of the app (the latter may overlap with other indicators such as 3, or 4. These should be coded separately). May include barriers and facilitators (feasibility) to use of the app and/or smartphone such as it being easy to fit self-report requests into a routine, or forgetting to carry the phone. May also include whether or not users/providers/practitioners would recommend the app to others/patients.

None of the literature-based apps featured in the app stores and therefore no user ratings were available. User ratings were captured for public apps at the time of identification in App Annie, and on the date of data extraction from their respective app stores (31st October 2018). Average user ratings across both sources were high, 4.45 and 4.02 out of 5 respectively. Average ratings for free, compared to paid apps, were 4.50 compared to 4.25 respectively in App Annie, and 4.43 compared to 4.18 respectively in stores. Average ratings for apps that promoted PA compared to apps that promoted multiple behaviours were 4.5 and 4.43 in App Annie and 4.55 and 4.35 respectively the app stores. Runtastic HR had the least ratings, while most other apps had thousands or hundreds of thousands of ratings.

Evidence for 16 apps reported outcomes of assessments of acceptability. Twelve apps reported both positive and negative outcomes and four reported only positive or only negative outcomes (Motimate, One You, Runtastic and Clue). Features and functions deemed acceptable varied between apps, with community aspects, reminders, photos, tailored messages, tracking, feedback, goals, videos, points systems and knowledge components being cited in apps from the literature. Custom plans, goals, the ability to capture a range of activities, different tracking modes, historical data, music integration, lack of advertisements, trainer modes, auto-pause settings, real-time behaviour logging and comparisons to other users were cited in public apps as being acceptable or preferred. Some users found apps and their devices convenient or easy to integrate into their lives, requiring little time commitment (Haptivity, iN Touch, Fitbit, Up). Other users liked the approach the app took to behaviour change, be it structured (MAPS), displaying behavioural correlations to aid learning and reflection (Health Mashups), un-judgemental (Health Mashups) or supportive (Motimate). Some users said they would continue to use the app, or would recommend it (iN Touch, SIGMA, Fitbit, Up). Evidence describing public apps often cited user ratings, with Runkeeper, Runtastic, Runtastic PRO and Fitbit all having high user ratings reported prior to the present study.

Users queried the relevance of some app features, including app-communities (when social media apps were so prevalent) (Haptivity), repetitive content (MAPS) and the use of statements of completed activities, rather than recommendations for how to improve (Health Mashups). Numerous requests for additional features were made, but were predominantly localised to apps from the literature. They included but were not limited to the ability to set user-defined goals or goals in general and more frequent messages, reminders and points, social elements, more chart-based and longitudinal feedback (MAPS). More frequently occurring new information was desired as well as recommendations for improvements and the ability to capture surrounding context for behaviours (Health Mashups). An auto-save feature for data entry was requested (iN Touch), as were simpler options for food tracking and stratified measures of comparison to other users (Fitbit). Some users were concerned about losing devices or forgot to enter data due to the location of the app on the home screen (Health Mashups, iN Touch). Other practical concerns arose related to cost, internet connectivity, (iN Touch, Runkeeper, Fitbit) and two apps reported lower user ratings than for this present study (One You and Fitbit) (see table 19 for details).

Table 19 Acceptability of apps

App	Rating	Number of ratings	Additional evidence* (source)
Literature-based apps			
Haptivity (Not available for download) PA	NA	NA	Some users were keen on the community/social motivation element of the app. Others questioned the added value in relation to current social media that they engage with. The *'activity reminder'* function of the app was well received and appeared to be successful in encouraging activity. Some users showed signs of *'photo fatigue'*. Others thought that a text-based platform would be uninspiring, suggesting that the app might need photos, but for these somehow to be more interesting, through promoting specific types of content or through enhancements added to the photos. Users found the app simple to integrate into their lives.(Forster et al., 2017)

MAPS (Not available for download) PA	1	NA	NA	Although individuals with access to in-app goal setting responded well to the structured approach, they often noted a desire to set maintenance goals at a lower level of PA (i.e., prior to meeting public health recommendations). Similarly, although individuals in group 1 and 3 (see table 17) were unaware of the availability of the goal-setting module, they frequently asked for in-app goal setting across a number of feedback categories. Individuals in all conditions typically found the tailored messaging to be highly motivating, and asked for more frequent messages. There was a small group of individuals who responded poorly to the educational content, and this appeared to be driven by the introductory video, which reiterated content delivered in an orientation appointment.
	2			Positive responses to the tracking feature per conditions (see table 17): feedback integration (all conditions), generally motivating (conditions 2/4). Negative comments on tracking included: no social component (2/4). Positive responses to in-app feedback: supportive of self-monitoring (all), generally motivating (all), and reminder for PA (1/3). Negative comments on in-app feedback included: desire for more chart options (2/4) and desire for more longitudinal graphical feedback (all).

	3			Positive responses to in-app goals: progressive structure and generally motivating (2/4). Negative comments on in-app goals included them being overly narrow/restrictive - i.e. they wanted to set them lower than the recommendations (2/4). Favourite features were reported as: email/text messages (all), in-app goals (2/4), in app tracking (all), and videos (1/3). Least favourite features were reported as: lack of longitudinal graphical feedback (2/4), need to track activities daily (2/4), videos (all - where videos deemed too similar), lack of in-app goals (1/3), email/text messages (1/3). Users wanted: daily reminders and feedback (all), in-app planning calendar (2/4), integration with activity monitors (all), in-app goal setting (1/3). Points system groups were positive about the system in terms of: motivated increased activity and they enjoyed the feeling of progression. Points groups wanted specific point values, more frequent levels/badges, material rewards. All groups (see table 17) found the knowledge component enjoyable but users wanted more broad health information. Goal booklet was positively received as practical, perceived as providing structure and fostering realistic goals (1/3) but booklet was easy to forget/misplace.(Fanning et al., 2017)
	4			
Unnamed app (Not available for download) PA+		NA	NA	NA
ATHENA apps		NA	NA	Assessed but not reported.

(Not available for download) PA+			
Health Mashups (Not available for download) PA+	NA	NA	**Pilot outcomes** Users liked that the correlations were made explicit in the daily feed. Most participants found interesting correlations and enjoyed this new way to learn more about their own wellbeing and appreciated the easy-to-understand, natural language presentation. Some participants forgot about the widget [app] or sensors. This was mostly due to users placing the widget on a secondary home screen that was not often displayed due to the widget's size. Most of the same observations remained significant from day to day, discouraging frequent use. Many participants reported wanting something *"new."* Users frequently reported not weighing themselves regularly as they did not want to see their weight after a day of heavy eating or sedentary activity. Reported not wanting to wear Fitbit on certain occasions or with certain clothes. Users wanted recommendations of what they could do to improve on identified problem areas. Users wanted raw sensor data so they could see latest progress towards goals. **Trial outcomes** Power of logging to increase reflection each day for understanding self, making sense and confirming user's perceptions. Some users wanted *"obvious"* activity statements rephrased to recommendations/goals e.g. *"Try to walk more on Tuesdays"*. However, others appreciated that the system is not *"judging"* and did not *"tell you what to do."* Some participants noted that it didn't capture context surrounding their

			behaviour. (Bentley et al., 2013)
iN Touch (Not available for download) PA+	NA	NA	Participants liked the ability to measure/track different things, not focused just on weight loss. Participants felt that the iPod was convenient and portable. A few were worried about the device being lost or stolen and reported that they left it at home when they went to certain events. This sometimes impeded their ability to track ODLs since they had to try and remember what they had done and enter it later. Tracking was not a big time commitment. The approximate time spent per day was reported to be 5–10 min. One commonly recommended improvement was an auto-save feature for data entry. Barriers to tracking were lack of regular access to Wi-Fi and inconsistent tracking due to other life priorities (e.g. motherhood). Although Wi-Fi might be available at certain locations, participants were not willing to go out of their way to connect. Participants indicated that they had not made ODL tracking a consistent habit. Others said that a major schedule change such as the transition from school year to summer disrupted their routine. Still others found that unavoidable life events challenged them (e.g. funerals). A few participants indicated they would continue to use iN Touch for tracking ODLs even after the study was over. One reported that she would continue using the iN Touch application because it *“is like a guide for me.”* (Kim et al., 2015)
Motimate (Not available for download) PA+	NA	NA	Users of the full version of the [Motimate] app reported that they felt more supported than those with the control app (Brindal et al., 2016)
Ngala	NA	NA	NA

(Not available for download) PA+			
SIGMA (Not available for download) PA+	NA	NA	**Pilot outcomes** There was a general satisfaction with the app. Several improvement suggestions were made. Pleasantness of the interaction scored 71.4% **Design study outcomes** 100% of the users said that they would recommend the SIGMA app and 85.8% of them declared themselves as generally satisfied and highly satisfied with the app. (Podina et al., 2017)
Descriptives:	**NA**	**NA**	**9/12 apps had acceptability outcomes reported.**
Public apps			
One You (Downloaded) PA	AA - 4.6 GP - 4.6	AA - 5596 GP - 11352	Average rating: 1.8/2.4 iOS, GP (Bondaronek et al., 2018)
Runkeeper (Downloaded) PA	AA - 4.5 GP - 4.5	AA - 508,136 GP - 525,892	Described as pros: Personal trainer building a custom plan, using prebuilt plans and possibility to define goals. It is able to record different sport activities. GPS mode and Stopwatch mode. Historical recording. Music player integrated. Free of advertising. Described as cons: Premium subscription is expensive. (Martinez-Nicolas et al., 2017) Popularity of app: 12 (frequency of downloads compared to 39 other physical activity and dietary apps that were evaluated). Average star rating in app store from 1-5: 4.5 (Direito et al., 2014).
Runtastic HR* (Downloaded) PA+	AA - 4.0 AS - 3.9	AA – 363 AS - 141	NA
Runtastic	AA - 4.5	AA - 843,884	Average user star rating was 4.5 in both iTunes and GP stores (Bondaronek

(Downloaded) PA+	GP - 4.5	GP - 887,875	et al., 2018)
Runtastic PRO (Downloaded) PA+	AA – 4.5 GP - 4.5	AA – 215,025 GP – 219,317	Described as pros: Personal trainer mode. Can be used in many different outdoor or indoor activities such as running, cycling, playing tennis, basketball, skiing or even on a treadmill. Free of advertising. Auto-pause. Described as cons: Robotic voice. Android version not as easy to navigate as iPhone version (Antón and Rodríguez, 2016) Average user star rating 4.6/4.5 iTunes, GP (Bondaronek et al., 2018)
Fitbit (Downloaded) PA+	AA - 3.9 GP – 3.9	AA - 374,196 GP - 473,221	Fitbit was recommended by 312 (43%) of a sample of responding clinicians asked about use of health apps with patients (Karduck and Chapman-Novakofski, 2018). Users reported it was acceptable to track food, using a 5 point Likert scale - 2.92 mean, 0.79 SD, 17% (n=2/12) agreed it's helpful for this and the perceptions about the acceptability of the food logging varied. Some reported logging being tedious and difficult, particularly if using a flip phone. Smart phone users liked the convenience of logging in real time but desired a *"simpler"* option for tracking their food intake, as there were no options for home-made foods on the food log menu, mainly restaurant options. Only one user reported accessing the app to log his food intake after 6 weeks. Time (peak planting season [users were farmers]) and difficulty with locating exact food-portion sizes were reasons the men cited on the survey for discontinuing logging. The men liked comparing their physical activity anonymously with other participants via the pseudonym identifier on the companion app: *"I think the motivation of just bein [sic] a part of a group makes you wanna [sic] do healthier things."* Users desired stratified measures, such as stride length and

			daily calorie intake so they could compare across other group members to provide relevance to their self-monitoring. Identity of each group member was also desired. Single-computer households caused competing demands among family members for Internet access, limiting ability to access apps. (Eisenhauer et al., 2017) App 'liked' and perceived as integrating well into user life. (Tong et al., 2018) Average user star rating: 3.9 iTunes, 4.0 GP (Bondaronek et al., 2018) Average user star rating: 3.0 (Pagoto et al., 2013)
Keep (Downloaded) PA+	AA - 5.0 AS - 4.8	AA – 366 AS - 8100	NA
Clue (Downloaded) PA+	AA - 4.8 GP – 4.8	AA - 610,548 GP - 702,616	Achieved high APPLICATIONS score: 13/15 showing presence of most of desired characteristics - absence of in-app purchases and adverts, connectivity (i.e. internet was not required to work), Android version was additionally available, Spanish language, custom reminder, track flow amount, track symptoms, track intercourse, alert for next menses, alert for fertility, average cycle length. Did not include fertility medications or professional involvement. (Moglia et al., 2016)
Samsung Health (Downloaded) PA+	AA - 4.3 GP - 4.3	AA - 544,136 GP - 728,608	NA
Up (Available version not functioning)	Version not reported in paper (But Up	Version not reported in paper (But Up	Responses to acceptability questions were broken down to show responses by participants under the age of 60 years as compared with the participants 60 years or older. All but one of the questions received a mean rating over 4

PA+	in GP on 26/6/19, 4.1)	in GP on 26/6/19, 63,764)	out of a possible 5 across both age groups, with only one receiving a 3.9 for the participants aged 60 or above (*"would you continue to wear the monitor?")* Across ages: Convenient to use app: Mean 4.74 (SD: 0.56).Would like to continue to use the app: Mean 4.53 (SD:0.84) Feasibility and acceptability findings showed that participants overall were compliant and reported enjoying the intervention. (Lyons et al., 2014)
Descriptives:	**AA rating range: 4.0-5.0** **Mean: 4.45** **App store rating range: 4.0-4.8** **Mean:4.02**	**AA rater range: 363-843,884** **Mean: 344,694.4** **App store rater range:141-887,875** **Mean:333,175**	**7/10 apps had acceptability assessed.**
TOTAL	**AA rating range: 4.0-5.0** **Mean: 4.45** **App store rating range: 4.0-4.8**	**AA rater range: 363-843,884** **Mean: 344,694.4** **App store rater**	**16/22 apps had had acceptability assessed.**

	Mean:4.02	**range:141-887,875** **Mean:333,175**	

AA = App Annie, AS = Apple app store, GP = Google Play app store, GPS = Global Positioning System, NA = Not applicable; NR = Not reported, ODLs = Observations of Daily Living, PA = Physical activity/physical activity targeted by apps, PA+ = physical activity and other behaviours targeted by apps, SD = Standard Deviation

*Runtastic HR appears to have lost user ratings between identification through AA and data extraction. This is unintuitive, however no explanation was apparent on closer inspection of the app store.

*Where possible, text is verbatim or reduced for length verbatim, or summarised using key terms from paper

5.4.2 Aesthetics

Quality indicator	Summary definition
Aesthetics	Visual attractiveness of the app interface design in terms of colours, fonts, and layout. How professional the design is. How pleasing to eye the design and layout is. Can include relevance of design to the behaviour.

Public and literature-based apps mean MARS scores for aesthetics were 4.53 and 3.24 respectively. Apps that promoted PA alone scored 3.81, compared to apps that promoted multiple behaviours, which scored 3.62. All except two literature-based apps (Unnamed, ATHENA), individually scored above average for aesthetics (see table 20).

Four apps reported outcomes for aesthetics (see appendix 18 for additional evidence that informed MARS scores). Little detail was provided, but colours, embedded maps, and attractive, elegant and sleek, inviting, designs were cited. Users also cited a preference for visualisation of data and progress for one app (Fitbit). Negative feedback was received for one literature-based app (SIGMA), relating to increasing brightness, an even text display and the need for a consistent design.

Table 20 Aesthetics of apps

App		Layout	Graphics	Visual appeal	Mean score
Literature-based apps					
Haptivity (Not available for download) PA		5	5	3	4.33
MAPS (Not available for download) PA	1	3	4	3	3.33
	2	4	3	3	3.33
	3	3	4	3	3.33
	4	4	3	3	3.33
Unnamed app (Not available for download) PA+		2	2	2	2.00
ATHENA apps (Not available for download) PA+		3	2	2	2.33
Health Mashups (Not available for download) PA+		4	4	3	3.67
iN Touch (Not available for download) PA+		3	4	2	3.00
Motimate (Not available for download) PA+		4	4	3	3.67
Ngala Healthy You, Healthy Baby (Not available for download) PA+		NR	NR	NR	NA
SIGMA (Not available for download) PA+		3	3	4	3.33
Mean		**3.45**	**3.45**	**2.82**	**3.24**
Median		**3**	**4**	**3**	
Public apps					

One You (Downloaded) PA	5	4	4	4.33
Runkeeper (Downloaded) PA	4	5	5	4.67
Runtastic HR (Downloaded) PA+	4	5	5	4.67
Runtastic (Downloaded) PA+	4	5	4	4.33
Runtastic PRO (Downloaded) PA+	4	5	4	4.33
Fitbit (Downloaded) PA+	5	5	5	5.00
Keep (Downloaded) PA+	4	5	5	4.67
Clue (Downloaded) PA+	3	5	5	4.33
Samsung Health (Downloaded) PA+	4	5	5	4.67
Up (Available version not functioning) PA+	4	4	5	4.33
Mean	**4.10**	**4.80**	**4.70**	**4.53**
TOTAL Mean	**3.76**	**4.10**	**3.71**	**3.86**
Median	**4**	**5**	**5**	
TOTAL Median	**4**	**4**	**3.5**	

NA = Not applicable; NR = Not reported, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.3 Physical activity (PA) measurement tool

Quality indicator	Summary definition
Behaviour measurement tools (e.g. PA measurement)	The type of tool/method used by the app to measure physical activity (or other behaviours as applicable) and its related validity and reliability. For example, objective or subjective measures. Also can include the definitive accuracy of the PA tools for capturing physical activity or perceived accuracy of in-app calculations.

Three apps did not actively capture PA (Haptivity, Runtastic HR PRO, One You), although the latter kept track of completed training sessions, users could let the sessions run without completing the session itself. Public apps tended to use both self-report and objective measures of PA, whereas literature-based apps tended to use self-report measures. The method of capture didn't seem to differ between apps that focused on PA alone, compared to those that promoted multiple behaviours – both used self-report and/or objective measures (see table 21). Thirteen apps reported using self-report measures to capture PA, however none appeared to be validated measures. Eleven apps used objective measures, most using inbuilt phone sensors such as GPS, accelerometers and pedometers (Unnamed, ATHENA, SIGMA, Runkeeper, Runtastic, Runtastic PRO, Fitbit, Keep, and Samsung Health). Some used linked wearables as well, or instead of, inbuilt sensors (ATHENA, Health Mashups, Up).

Eight apps reported justification or evidence for their choice of PA measure, or why they didn't include one (Haptivity), with evidence more forthcoming for literature-based apps. The various objective and subjective measures of PA are subject to well acknowledged strengths and weaknesses (Sylvia et al., 2015; Taylor, 2014). Additional strengths were reported including user preference for in-app tracking (MAPS), use of objective measures that weren't resource intensive (Unnamed), methods that enhanced accurate or valid data capture (Unnamed, ATHENA, Fitbit, Keep, Samsung Health) or increased the variety of data capture that was possible (ATHENA, Motimate, Runkeeper) and easy, low effort, brief, or automatic methods of capture (Health Mashups, iN Touch, Motimate, SIGMA, Keep, Clue, Up).

Weaknesses included the lack of activity capture (Haptivity, One You, Runtastic HR), a lack of different types of activity tracking or integration with wearables (MAPS), complicated and difficult to understand activity tracking (MAPS, Motimate), the inability to track a variety of activities or activities in more varied settings (Unnamed app, Runkeeper, Runtastic, Runtastic PRO, Fitbit, Keep, Clue, Samsung Health), resource heavy methods or those that required internet or GPS signal (Unnamed, ATHENA, iN Touch, Runkeeper, Runtastic, Runtastic PRO, Fitbit), technology challenges (Health Mashups, Keep), measures without a reported evidence base (Health Mashups, Motimate) and a questionable or lack of reliable or valid objective measure (Runkeeper, Runtastic, Runtastic PRO, Fitbit, Keep, Up) (see appendix 19.)

Measures of PA also revealed the characteristics of PA that were captured/promoted, with apps varying as to whether duration, frequency and intensity were included (see table 21).

Table 21 Physical activity measurement tools used in apps

<table>
<tr><th colspan="2">App</th><th>Method of PA measurement</th></tr>
<tr><td colspan="3">Literature-based apps</td></tr>
<tr><td colspan="2">Haptivity
(Not available for download)
PA</td><td>PA not captured</td></tr>
<tr><td rowspan="4">MAPS
(Not available for download)
PA</td><td>1</td><td rowspan="4">Self-monitoring - entered the activity for conditions 1/3, but in conditions 2/4 (see table 17) user could specify between aerobic or non-aerobic goal activities using quick-track buttons and duration and intensity (moderate, hard etc.) using slider buttons.</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td colspan="2">Unnamed app
(Not available for download)
PA+</td><td>Location tracking via GPS, network provider, and passive provider, each suitable for different activity conditions. Internal sensors also used including altimeter and accelerometer. Exercise, distance travelled and speed captured.

Using information from location data, the app calculates the main parameters concerning a healthy behaviour, such as total travelled distance, average speed and time spent. In addition the app shows the users route using maps including Google Maps API and OSMDroid to display OpenStreetMap (OSM).</td></tr>
<tr><td colspan="2">ATHENA apps
(Not available for download)
PA+</td><td>“Activity recognition recognizes the physical activities of the subjects through the use of the embedded sensors of a smartphone. Authors developed an app “Action Logger” to recognise outdoor activities by utilising the accelerometer, mic, GPS and WiFi sensor. It is position free and energy aware, to activate the sensors when required. Activities, such as walking, jogging, riding on a bus or taking a subway and staying in place can be accommodated. Also included a specific step counter app.” (Fahim et al., 2014)

When there's a change in activity to the user standing still, the user is prompted to label the previous activity as walking, standing still, and running or on bus. The raw data includes embedded sensor data, activity label, and start and end of activity. It detects an activity based on three seconds (150 samples) of data.(Cleland et al., 2013)

In addition, a Daily Activity Monitor which captured PA via GPS, compared activities to the prescribed activities to determine discrepancies. Distance and time spent performing the activities were captured. (Saleem et al., 2012)</td></tr>
</table>

	Wearables for specific sport types were also used, but as these were in addition to PA data captured by the app and not the main focus, they will not be reported here (Fahim et al., 2014).
Health Mashups (Not available for download) PA+	**Pilot measures** Objectively captured using Fitbit wearable device to capture step count every day (version unspecified, but image in pilot paper suggests Fitbit Ultra). Worn clipped to trousers. Self-reported amount of exercise captured in app using 5 star scale item *"Exercise: How have you been exercising today?*" where 1=none, 3=normal, 5=more. Users encouraged to log daily. **Trial measures** Objectively captured using Fitbit wearable device (unspecified) to capture step count. Worn clipped to trousers.
iN Touch (Not available for download) PA+	Self-reported by users who could add any activity and its duration.
Motimate (Not available for download) PA+	Self-reported by user. Single item question asks users *'What exercise have you done today?'* User can respond with as many entries as desired, describing the activity, duration and intensity.
Ngala (Not available for download) PA+	Not adequately reported, just stated as '*API*
SIGMA (Not available for download) PA+	Objectively captured steps per day by inbuilt smartphone pedometer
Descriptives:	**7/12 self-report measures** **2/12 objective measures** **2/12 use both**
Public apps	
One You (Downloaded) PA	PA not captured other than to mark a session as complete.
Runkeeper (Downloaded) PA	Objectively measured using GPS tracking when activity is outside. Otherwise, time, distance, pace of session is captured objectively using stop-watch function. Training sessions (My first 5k) are also captured when completed. Activity can also be manually logged.
Runtastic HR (Downloaded)	PA not captured.

PA+	
Runtastic (Downloaded) PA+	Objectively measures duration, distance, calories and average pace min/km using GPS where applicable. Can also add activity manually - type, duration, distance, calories, date, time as applicable. Can also add wearable.
Runtastic PRO (Downloaded) PA+	App captures duration, distance, calories and average pace min/km using GPS where applicable. Can also add activity manually - type, duration, distance, calories, date, time, HR as applicable. Can also add wearable.
Fitbit (Downloaded) PA+	Objectively measures step count using internal phone sensor, or will track activities in real time via GPS. Can also add activities manually.
Keep (Downloaded) PA+	Objectively measures step count using internal pedometer. Can also add activities manually. Completed in-app training plans will also automatically be recorded/added to activity log.
Clue (Downloaded) PA+	Self-report. Exercise can be logged during daily data entry by clicking an icon for either running, yoga, biking or swimming. No other options are allowed, but the information button for this item shows that each item includes variations of these behaviours e.g. running can include jogging, sprinting or speed-walking.
Samsung Health (Downloaded) PA+	Objectively measures step count using internal pedometer. Workouts also detected automatically by app using GPS e.g. running map pops up. Can also add activities manually.
Up (Available version not functioning) PA+	Objectively measured using UpBand commercial accelerometer and the app providing information on physical activity (total number of steps, active time, longest active time, longest idle time, and exercise intensity).
Descriptives	**1/10 self-report measures** **1/10 objective measures** **6/10 use both**
TOTAL	**8/22 self-report measures** **3/22 objective measures** **8/22 use both** **3/22 did not capture PA**

API = Application Program Interface, GPS = Global Positioning System, HR = Heart Rate, PA = Physical activity/Physical activity targeted by app, PA+ = Physical activity and other behaviours targeted by app

5.4.4 Credibility/trustworthiness/usefulness

Quality indicator	Summary definition
Credible, trustworthy/appropriate or useful/essential information	Content of the app is likely to be accurate or believable – not making impossible or implausible claims. Content is safe for users, won't harm them or will minimise harm, or provides a caveat for medical information that requires seeing a professional. Information/app appears useful.

Public and literature-based apps had mean credibility scores of 3.91 and 3.53 respectively. Apps that focused on PA had a mean credibility score of 3.71 compared to apps targeting multiple behaviours (3.70). One literature-based app (Unnamed) scored less than average for credibility (see table 22 and appendix 18 for additional evidence that informed MARS scores).

Four apps reported only positive outcomes for credibility (ATHENA, SIGMA, Runtastic PRO, Up), citing accurate recommendations based on entered data, the ability to learn from content and resist food temptations, that the app was reliable, or that included content such as step goals was useful and information was credible and relevant. Apps were perceived as including informative content or feedback (MAPS, Health Mashups, iN Touch and Fitbit). However, some literature-based apps were still seen as lacking useful or including redundant content (Haptivity, MAPS, iN Touch) (see appendix 18).

Users of one app perceived it as truthful and scientific, but also as reporting questionable or contradictory feedback (Health Mashups). Similarly, public apps had mixed evidence with respect to their accuracy or validity when tested or compared to other apps/measures. Runtastic HR, Fitbit, and Samsung Health were described as performing well or poorly depending on the test. Others also performed poorly in tests of validation and accuracy (Runkeeper, Runtastic, and Clue) (see appendix 18).

Table 22 Credibility of apps (from Information section of MARS)

App		Accuracy	Goals	Quality of information	Quantity of information	Visual information	Credibility	Evidence base	Mean score
Literature-based apps									
Haptivity (Not available for download) PA		NR	3	4	1	NA	5	2	3.00
MAPS (Not available for download) PA	1	NR	3	4	4	3	4	2	3.33
	2	NR	4	4	4	4	4	2	3.67
	3	NR	4	4	4	3	4	4	3.83
	4	NR	5	4	4	3	4	4	4.00
Unnamed app		NR	2	2	1	3	4	NA	2.40

(Not available for download) PA+								
ATHENA apps (Not available for download) PA+	NA	2	3	3	4	4	NA	3.20
Health Mashups (Not available for download) PA+	NR	3	4	4	4	3	2	3.33
iN Touch (Not available for download) PA+	NR	3	3	1	4	4	3	3.00
Motimate (Not available for download) PA+	NR	4	5	2	4	4	4	3.83
Ngala (Not available for	NR	NR	5	NR	NR	4	NA	4.50

download) PA+								
SIGMA (Not available for download) PA+	NA	5	4	5	NR	5	2	4.20
Mean	**NA**	**3.45**	**3.83**	**3.00**	**3.56**	**4.08**	**2.78**	**3.53**
Median	**NA**	**3**	**4**	**4**	**4**	**4**	**2**	
Public apps								
One You (Downloaded) PA	5	5	5	5	2	5	NA	4.50
Runkeeper (Downloaded) PA	4	5	4	1	5	3	NA	3.67
Runtastic HR (Downloaded) PA+	5	5	3	1	4	3	NA	3.50
Runtastic (Downloaded) PA+	4	4	3	2	5	3	2	3.29
Runtastic PRO	5	4	3	2	5	3	NA	3.67

(Downloaded) PA+								
Fitbit (Downloaded) PA+	4	5	5	4	4	3	2	3.86
Keep (Downloaded) PA+	5	5	5	4	5	3	NA	4.50
Clue (Downloaded) PA+	5	NA	4	5	3	4	2	3.83
Samsung Health (Downloaded) PA+	5	5	4	5	4	3	NA	4.33
Up (Available version not functioning) PA+	Version unclear – can't code	5	5	4	5	4	1	4.00
Mean	**4.67**	**4.78**	**4.10**	**3.30**	**4.20**	**3.40**	**1.75**	**3.91**
TOTAL Mean	**4.67**	**4.05**	**3.95**	**3.14**	**3.89**	**3.77**	**2.46**	**3.70**
Median	**5**	**5**	**4**	**4**	**4.5**	**3**	**2**	

TOTAL Median	**5**	**4**	**4**	**4**	**4**	**4**	**2**	

NA = Not applicable; NR = Not reported, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.5 Currency and maintenance

Quality indicator	Summary definition
Currency/Maintenance of the app and its documentation	The date of the last update/regularly updated (no consensus, range between 1 and 6 months since last update) and date of creation and last update are reported. App documentation is updated as well as the app itself.

Literature-based apps tended not to report release or update dates, with two giving minimal details (Health Mashups, Ngala). Public apps reported both, with one being discontinued (Up). Eight of the public apps were updated within a month of data extraction, suggesting regular maintenance. One app had been updated within the last three months (Keep) (see table 23).

Table 23 Currency and maintenance of apps

App	Original release Date	Last update	Date data extraction performed
Literature-based apps			
Haptivity (Not available for download) PA	NR	NR	NA
MAPS (Not available for download) PA 1 2 3 4	NR	NR	NA
Unnamed app (Not available for download) PA+	NR	NR	NA
ATHENA apps (Not available for download) PA+	NR	NR	NA
Health Mashups (Not available for download) PA+	Not reported – but piloted 2011	Not reported – but trial version created	NA
iN Touch (Not available for download) PA+	NR	NR	NA
Motimate (Not available for download) PA+	NR	NR	NA
Ngala (Not available for download) PA+	June 2012	NR	NA
SIGMA (Not available for download) PA+	NR	NR	NA
Descriptives:	**1/12 apps reported release date**	**0 apps reported update date**	**NA**
Public apps			
One You (Downloaded) PA	05/03/2019	20/05/2019	04/06/2019
Runkeeper (Downloaded) PA	29/12/2011	04/06/2019	13/06/2019
Runtastic HR (Downloaded) PA+	20/12/2012	06/06/2019	07/06/2019
Runtastic (Downloaded)	24/01/2012	03/06/2019	06/06/2019

PA+			
Runtastic PRO (Downloaded) PA+	24/01/2012	03/06/2019	08/06/2019
Fitbit (Downloaded) PA+	16/03/2012	23/05/2019	11/06/2019
Keep (Downloaded) PA+	08/12/2017	14/03/2019	09/06/2019
Clue (Downloaded) PA+	08/10/2014	04/06/2019	05/06/2019
Samsung Health (Downloaded) PA+	06/04/2015	14/06/2019	16/06/2019
Up (Available version not functioning) PA+	Not captured (version unknown)	Discontinued (UP in GooglePlay – 2/10/17)	22/05/2019
Descriptives:	**9/10 apps reported release date Range: 29/12/2011 – 05/03/2019**	**10/10 apps reported last update date Range: Discontinued – 14/06/19**	**8/10 apps were updated within one month of data extraction**
TOTAL	**10/22 apps reported release date**	**10/22 apps reported last update date**	**8/22 apps were updated within one month of data extraction**

NA = Not applicable; NR = Not reported, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.6 Development process and team

Quality indicator	Summary definition
Development process and teams	Affiliations or credentials of app development team (University, Industry, Government, Commercial or Non-commercial etc.), involvement of experts and users in development.

More literature-based apps explicitly reported involving potential users in app development or piloting than public apps (five and two respectively). However, all public apps had user-reviews available to them at all times (see 5.4.15), which could be used for development purposes after the initial version had been launched. One developer, for three included apps, explicitly reported this on their website (Runtastic) (see table 24 for further details of user involvement).

Most apps (12/22) had commercial affiliations, especially the public apps. Three had government affiliations (Motimate, Ngala, One You). Public apps were from established companies and thus included experts in technology in their development teams, but literature-based apps more frequently reported the involvement of health or behaviour change experts.

Only public apps reported evidence on their development teams/processes. Four apps were reported as explicitly not including experts (behavioural experts as well as users) in the development process (One You, Runtastic, Runtastic PRO, Fitbit). One reported health professional involvement, but it was unclear in what capacity – either during development, or within the app itself (Clue).

Table 24 App development processes and team

App		Inclusion of target groups in development	Developer and affiliations	Expertise in development team including authors	Additional evidence
Literature-based apps					
Haptivity (Not available for download) PA		Iterative focus groups with users for co-design of a prototype, based on original specification by researchers. Discussed app needs and wants, functionality preferences, barriers to PA and feasible PA. Users tested app and discussed acceptability, appropriateness, aesthetics and functionality of prototype. Four rounds of development, each with feedback from users. Users completed a process evaluation of prototype and preliminary efficacy testing of prototype. Demographics same as sample.	Authors and app designer. University, commercial.	Behavioural science, health researchers, specialist app designer, developers and psychologists.	NA
MAPS (Not available for download)	1	None reported	Not reported – authors. University.	Specialists from physical activity, health, psychiatry/ neurobehavioural sciences.	NA
	2				
	3				

PA 4				
Unnamed app (Not available for download) PA+	None reported.	Not reported – authors University.	Health and sport psychology, electronic engineering.	NA
ATHENA apps (Not available for download) PA+	None reported.	Authors. University, Government.	Extensive technology and computer engineering expertise.	NA
Health Mashups (Not available for download) PA+	Users tested a pilot version before the full trial. Version based on suggestions for changes that emerged from qualitative data and quantitative usage data. Participants same as study group except residents of Stockholm as well as USA.	Authors including Motorola. Commercial, University.	Research and technology experts. Apps and smartphone expertise.	NA
iN Touch (Not available for download) PA+	Participatory design approach used. Youth advisory board (YAB) suggested refinements to the technology, testing authors initial design assumptions and suggesting how to adapt it for users' contexts and environments. YAB came up with 14 design recommendations related to flexibility and individual control, a non-judgemental	Collaborative development including technology partner TheCarrot. University, commercial. Also	Health and communication techniques. Medical and clinical expertise including health and mood. Behavioural and emotional technology trackers. Research and recruitment. App design.	NA

	approach, context sharing, tech support and data entry requirements to inform the design. YAB performed initial user testing before larger scale testing conducted. User testing sessions involved a user scenario and users being observed through their step-by-step use of the app and followed up with an interview for feedback about preferences, suggestions for changes, ease of use, feasibility. Same as sample.	non-profit organisation via Project HealthDesign.		
Motimate (Not available for download) PA+	Nothing reported.	Authors. Government.	Target behaviour expertise.	NA
Ngala (Not available for download) PA+	Focus groups and interviews with potential users revealed maternal use of online devices and resources for lifestyle information during pregnancy and early postnatal months. Identification of main online resources used to seek information, usefulness of online information, amount/format of information required, source of online information sought, trustworthiness of online information, gaps in useful forms of online information and preferred formats for presentation took place. Preferences from users were converted into	Authors. University, government, non-profit organisation charity.	Exercise, health science and promotion. Child health.	NA

	a gap analysis scale which was applied to existing online resources. The results from the qualitative work and gap analysis was fed back to government and non-government stakeholders and health practitioners, to allow them to decide on the online tools to develop. Participants as reported in sample.			
SIGMA (Not available for download) PA+	A sample of 21 Caucasian volunteers (4 male, 17 female) were enrolled in a pilot study to test the app. Participants ranged in age from 20 to 51 (M= 27.38, SD= 7.83). The majority of the participants were employed (N = 13) and almost half of the participants had at least a bachelor's degree (N = 10). Participants were normal weight adults (N = 12; BMI = 18.5 – 24.9 kg/m2), as well as overweight individuals (N = 9; BMI = 25-29.9kg/m2). The participants enrolled in the study were provided with the SIGMA app for the duration of one week and asked to complete specific tasks, as well as completing a System Usability Scale, Player Experience Scale - 16, SIGMA questionnaire, and an open-ended section with suggestions and comments regarding the app (Podina et al., 2018)	Authors. University.	Psychology, Psychotherapy, Evidence-based psychological assessment and interventions, computer science.	NA

Descriptives	5/12 apps involved potential users in development or piloting of the app	3/12 apps had commercial developers as part of the team 2/12 had government affiliations	6/12 app development teams included technology expertise 10/12 app development teams included health or behaviour change expertise	No feedback on development team/process reported
Public apps				
One You (Downloaded) PA	Not reported.	Public Health England. Government.	Public Health Promotion.	Development process: No experts involved. Only one to report a non-commercial affiliation (Bondaronek et al., 2018)
Runkeeper (Downloaded) PA	Not reported.	Runkeeper (Owned by ASICS) Commercial.	App developers.	NA
Runtastic HR (Downloaded) PA+	Not reported.	Runtastic. Commercial.	App developers.	NA
Runtastic (Downloade	Not reported.	Runtastic.	App developers.	Development process: No

d) PA+		Commercial.		expert involvement (Bondaronek et al., 2018)
Runtastic PRO (Downloaded) PA+	Not reported.	Runtastic. Commercial.	App developers.	Development: No expert involvement (Bondaronek et al., 2018)
Fitbit (Downloaded) PA+	Not reported.	Fitbit, Inc. Commercial.	App developers.	Development: No expert involvement. (Bondaronek et al., 2018)
Keep (Downloaded) PA+	Not reported.	Keep Inc. Commercial.	App developers.	NA
Clue (Downloaded) PA+	Website suggests app development includes input from users.	BioWink GmbH. Commercial.	App developers, including data and research scientists. Collaborations with top academic institutions and Medical Board of leading specialists.	Development: health professional involvement (Moglia et al., 2016)
Samsung Health (Downloaded) PA+	Not reported.	Samsung. Commercial.	App developers.	NA

Up (Available version not functioning) PA+	RCT pilot trial of the whole intervention with a focus on feasibility and acceptability of the app and monitor. Participants had BMIs of 25-35 - average 30.3 (SD 3.5), <60min activity per week. Major exclusion criteria included self-reported habitual physical activity more than 60 min per week, health issues that might preclude safe walking, psychological issues that might interfere with full participation, current use of a wearable electronic activity monitoring system, and endorsing cardiovascular risk questions on the Physical Activity Readiness Questionnaire. If the only questions endorsed had to do with taking medication, individuals could participate if they provided a doctor's consent. Weight average 82.36 (SD 10.81). 55-79 years old 61.48 years old (SD 5.60). 85% (34/40) female. 27 (68%) college degree. 65% (26/40) white, 13% black, 15% other, 28% Hispanic. (Lyons et al., 2017)	Jawbone. Study - University. App - Commercial.	Commercial app and wearable activity tracker developers.	NA
Descriptives	**2/10 apps reported involving potential users in development or piloting**	**9/10 apps had commercial affiliations** **1/10 apps had government affiliations.**	**9/10 app development teams included technology expertise** **2/10 app development teams included health or behaviour change expertise**	**5/10 reported feedback on development process/team.** **4/10 apps explicitly reported that no**

				experts or users, were involved in app development
TOTAL	**7/22 apps involved potential users in development or piloting of the app**	**12/22 apps had commercial developers as part of the team** **3/22 had government affiliations**	**15/22 app development teams included technology expertise** **12/22 app development teams included health or behaviour change expertise**	**5/22 reported feedback on development process/team** **4/22 apps explicitly reported that no experts or users, were involved in app development**

BMI = Body Mass Index, M = Mean, NA = Not applicable, PA = Physical activity/Physical activity targeted by app, PA+ = Physical activity and other behaviours targeted by app, RCT = Randomised Control Trial, SD = Standard Deviation,

5.4.7 Effectiveness/potential impact

Quality indicator	Summary definition
Effectiveness/Potential impact	Evidence of improvements in relevant health outcomes/behaviours. Potential for impact on behaviour or health.

Literature-based and public apps reported mean MARS scores of 3.11 and 2.69 respectively. The latter was below the average score, but more literature-based apps scored below average individually (6 compared to 5 public apps). Apps that focused on PA scored 3.58 compared to apps that targeted multiple behaviours (2.61) (see table 25 and appendix 18 for additional evidence that informed MARS scores).

Three apps reported effectiveness outcomes, suggesting they or specific content were motivating or encouraging (Haptivity – reminders, Fitbit - comparisons) or influenced PA levels (Runkeeper –social aspect and goals) or being healthier (Fitbit) (see appendix 18).

Table 25 Effectiveness and potential impact of apps

App		Evidence base*	Awareness	Knowledge	Attitudes	Intention to change	Help seeking	Behaviour change	Mean score (including evidence base)
Literature-based apps									
Haptivity (Not available for download) PA		2	2	2	3	4	1	4	2.57
MAPS (Not available for download) PA	1	2	5	5	3	4	3	4	3.71
	2	2	5	5	3	5	3	3	3.71
	3	4	5	5	4	4	3	5	4.29
	4	4	5	5	4	5	3	5	4.43
Unnamed app (Not available for download) PA+		NA	1	2	2	3	1	2	1.83
ATHENA apps		NA	3	2	1	3	2	3	2.33

(Not available for download) PA+								
Health Mashups (Not available for download) PA+	2	4	3	3	3	3	3	3.00
iN Touch (Not available for download) PA+	3	3	3	2	3	2	3	2.71
Motimate (Not available for download) PA+	4	3	3	2	2	1	1	2.29
Ngala (Not available for download) PA+	NA	4	4	4	4	3	4	3.83
SIGMA (Not available for download)	2	3	3	2	3	2	3	2.57

PA+								
Mean	**2.78**	**3.58**	**3.50**	**2.75**	**3.58**	**2.25**	**3.33**	**3.11**
Median	**2**	**3.5**	**3**	**3**	**3.5**	**2.5**	**3**	
Public apps								
One You (Downloaded) PA	NA	4	4	4	3	4	4	3.83
Runkeeper (Downloaded) PA	NA	1	2	3	2	3	4	2.50
Runtastic HR (Downloaded) PA+	NA	1	1	1	1	1	1	1.00
Runtastic (Downloaded) PA+	2	1	2	2	3	2	4	2.29
Runtastic PRO (Downloaded) PA+	NA	1	2	2	3	2	4	2.33
Fitbit (Downloaded)	2	3	3	4	4	3	4	3.29

PA+								
Keep (Downloaded) PA+	NA	2	3	3	4	3	4	3.17
Clue (Downloaded) PA+	2	2	2	2	2	1	2	1.86
Samsung Health (Downloaded) PA+	NA	4	4	3	3	2	4	3.33
Up (Available version not functioning) PA+	1	5	5	5	2	3	2	3.29
Mean	**1.75**	**2.40**	**2.80**	**2.90**	**2.70**	**2.40**	**3.30**	**2.69**
TOTAL Mean	**2.46**	**3.05**	**3.18**	**2.82**	**3.18**	**2.32**	**3.32**	**2.92**
Median	**2**	**2**	**2.5**	**3**	**3**	**2.5**	**4**	
TOTAL Median	**2**	**3**	**3**	**3**	**3**	**2.5**	**4**	

NA = Not applicable, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

*Evidence Base was originally part of the Information section of MARS, but has been added to the Potential Impact MARS items here as it fits with the defined effectiveness criterion compiled for this thesis.

5.4.8 Engagement

Quality indicator	Summary definition
Engagement	Use of methods to encourage user interactivity with the app, can include use of certain strategies or features that promote/inhibit for example, feedback, tailoring, prompts/reminders, gamification.

Public and literature-based apps mean MARS scores for engagement were 3.74 and 3.07 respectively, with five of the latter scoring below average (Haptivity, MAPS 1, Unnamed, iN Touch, Ngala). Apps promoting only PA scored the same as apps promoting multiple behaviours (mean 3.37) (see table 26 and appendix 18 for additional evidence that informed MARS scores).

Seven apps reported engagement outcomes. Negative feedback included lengthy or repetitive content that discouraged use (MAPS, Health Mashups), while positive feedback included references to content, such as the ability to view long-term trends (Health Mashups), the variety of, and, optional functions (iN Touch), gamification (Runkeeper), achieving goals (Runtastic), user experience or design features (Fitbit) and content perceived as being specific to the user (Up) (see appendix 18).

Table 26 App engagement

App		Entertainment	Interest	Customisation	Interactivity	Target group	Mean score
Literature-based apps							
Haptivity (Not available for download) PA		3	2	2	3	3	2.60
MAPS (Not available for download) PA	1	2	3	2	3	3	2.60
	2	3	4	3	4	4	3.60
	3	4	4	2	4	3	3.40
	4	5	5	3	5	3	4.20
Unnamed app (Not available for download) PA+		2	2	1	2	3	2.00
ATHENA apps (Not available for download) PA+		2	4	4	4	4	3.60
Health Mashups		2	4	2	4	4	3.20

(Not available for download) PA+						
iN Touch (Not available for download) PA+	2	3	3	2	3	2.60
Motimate (Not available for download) PA+	2	4	2	4	4	3.20
Ngala (Not available for download) PA+	2	NR	NR	2	NR	2.00
SIGMA (Not available for download) PA+	5	4	1	4	5	3.80
Mean	**2.83**	**3.55**	**2.27**	**3.42**	**3.55**	**3.07**
Median	**2**	**4**	**2**	**4**	**3**	
Public apps						
One You (Downloaded) PA	2	3	3	2	3	2.60
Runkeeper (Downloaded) PA	5	4	5	5	4	4.60
Runtastic HR	2	2	2	2	2	2.00

(Downloaded) PA+						
Runtastic (Downloaded) PA+	4	3	4	5	3	3.80
Runtastic PRO (Downloaded) PA+	4	3	4	5	3	3.80
Fitbit (Downloaded) PA+	5	4	4	5	5	4.60
Keep (Downloaded) PA+	4	4	4	5	4	4.20
Clue (Downloaded) PA+	2	4	4	4	4	3.60
Samsung Health (Downloaded) PA+	5	5	4	5	4	4.60
Up (Available version not functioning) PA+	4	5	2	3	4	3.60
Mean	**3.70**	**3.70**	**3.60**	**4.10**	**3.60**	**3.74**
TOTAL Mean	**3.23**	**3.62**	**2.90**	**3.73**	**3.57**	**3.37**
Median	**4**	**4**	**4**	**5**	**4**	
TOTAL Median	**3**	**4**	**3**	**4**	**4**	

NR = Not reported, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.9 Evidence based content and components

Quality indicator	Summary definition
Evidence-based content/components	Use of techniques, strategies, information, practice or recommendations that are based on scientific evidence that demonstrates their usefulness. This can include behaviour change techniques or ‘predictors’ associated with improvements in behaviours in general, or the target behaviour, prescribed behaviours or practices advocated by reliable government bodies such as Public Health England, or the National Institute of Health. Techniques can be captured by referring to taxonomies such as the Behaviour Change Taxonomy v1 (Michie et al., 2013)

5.4.9.1 Presence of Behaviour Change Techniques (BCTs)

Public apps included more variety (n=36 compared to n=33) and more BCTs than literature-based apps in total and on average (mean 15.8, compared to a mean of 9.33). Both categories of apps used the most techniques from the Feedback and Monitoring cluster. Apps promoting multiple behaviours featured a mean of 12.13 BCTs, with apps promoting PA including a mean of 12.71. Motimate and SIGMA included BCTs that targeted diet and eating alone, as well as other BCTs that targeted PA (highlighted yellow in table 27). Ten apps included BCTs targeting PA that were delivered by non-app resources such as websites (all versions of MAPS, ATHENA, iN Touch, Runtastic HR PRO, Runtastic, Fitbit, Up).

Apps included between one and 12 BCTs that were associated with increased levels of PA. Public apps, on average, included more (mean 8.40) than literature-based apps (mean 3.67). Apps promoting PA included a mean of 5.71 BCTs and apps promoting multiple behaviours included 5.87.

One literature-based app, did not include a BCT with mixed-evidence for its association with changing levels of PA (Haptivity). The remaining apps included one or two and none included all three BCTs with mixed evidence.

One app included the BCT associated with decreases in PA (SIGMA) (see table 27 and appendix 20 for further details).

Table 27 Presence of BCTs in apps

	App																							
	Literature-based apps													**Public apps**										
		MAPS																						
	Haptivity	1	2	3	4	Unnamed	ATHENA	Health Mashups	iN Touch	Motimate	Ngala	SIGMA	Total	One You	Runkeeper	Runtastic HR	Runtastic	Runtastic PRO	Fitbit$	Keep	Clue	Samsung Health	Up*	Total
BCT	PA					PA+								PA		PA+								
1.1		+ +	+ +	+ +	+ +					+		+	6	+	++		++	++	++	++		++	++	8
1.2		+	+	+	+		+			++		+	7	++										1
1.3									++	+	+		3		++				++	++		++		4
1.4			+ +		+ +		++		+	++			5		++		++	++	++			++		5
1.5							+						1									++		1
1.6							+			+		+	3		++		+	+	++			++		5
1.7													0									+		1
1.8									+				1											0
1.9													0											0
SUM													**26**											**25**
2.1						++	+						2											0

2.2		+ +	+ +	+ +	+ +	++	++	++	++	++	+	+	11	++	++	++	++	++	++	++	++	++	++	10
2.3		+ +	+ +	+ +	+ +	++	++	++	++	++	+	++	11	+	++	++	++	++	++	++	++	++	++	10
2.4										++	++		2		++	++			++	++		++		5
2.5													0											0
2.6						++							1		+	++		+				+		4
2.7		+	+	+	+		+	++		+	+	+	9	+	++	+			++	++		++	+	7
SUM													**36**											**36**
3.1		+	+	+	+		+		++	++		++	8	++	++	+	+	+	+	+		+	++	9
3.2													0	+										1
3.3													0											0
SUM													**8**											**10**
4.1													0	++	+	++	+	+	++	++				7
4.2												+	1											0
4.3													0											0
4.4													0											0
SUM													**1**											**7**
5.1													0	++	+	+			+	++	++	++		7
5.2	++							++				+	3			+								1
5.3													0											0
5.4	++	+ +	+ +	+ +	+ +							++	6	++	++	+	++	++		++				6
5.5													0											0
5.6													0	+					++		++			3
SUM													**9**											**17**
6.1													0						++	++		++		3
6.2	++						+						2		++	++	++	++	++	++		++		7

6.3													0											0
SUM													**2**											**10**
7.1	++	+ +	+ +	+ +	+ +				++	++			7	++	++	++	+	+	++	++		++	+	9
7.2													0											0
7.3													0											0
7.4													0											0
7.5													0											0
7.6													0											0
7.7													0											0
7.8													0											0
SUM													**7**											**9**
8.1												+	1											0
8.2													0											0
8.3	++												1											0
8.4													0											0
8.5													0											0
8.6													0											0
8.7			+		+								2	++	+				+					3
SUM													**4**											**3**
9.1		+ +	+ +	+ +	+ +								4	++		+	+	+			+			5
9.2													0											0
9.3													0											0
SUM													**4**											**5**
10.1												+	1		++	+	+	+	+					5
10.2												+	1		++				++					2
10.3				+	+							+	3		++				++	+		++		4

10.4	++			+					+	3		++	++	++	++	++	++		++	++	8
10.5										0											0
10.6									+	1		+				+					2
10.7										0	+										1
10.8										0											0
10.9										0		+									1
10.10										0											0
10.11										0											0
SUM										**9**											**23**
11.1										0											0
11.2									+	1											0
11.3										0											0
11.4										0											0
SUM										**1**											**0**
12.1									+	1	+										1
12.2										0	+	+	+	+	+	++	+		++		8
12.3										0											0
12.4							+		+	2											0
12.5					++					1										+	1
12.6										0							+		+		2
SUM										**4**											**12**
13.1										0											0
13.2							++			1											0
13.3										0											0
13.4										0											0
13.5										0											0
SUM										**1**											**0**
14.1										0											0

14.2													0											0
14.3													0											0
14.4													0											0
14.5													0											0
14.6													0											0
14.7													0											0
14.8													0											0
14.9													0											0
14.10													0											0
SUM													**0**											**0**
15.1													0		+									1
15.2													0											0
15.3													0											0
15.4													0											0
SUM													**0**											**1**
16.1													0											0
16.2													0											0
16.3													0											0
SUM													**0**											**0**
SUM total	**6**	**9**	**11**	**10**	**12**	**4**	**11**	**5**	**7**	**13**	**5**	**19**	**112**	**17**	**24**	**16**	**14**	**15**	**22**	**17**	**5**	**20**	**8**	**158**
Mean	**9.33 BCTs**													**15.8 BCTs**										
TOTAL	**33 BCTs used**													**36 BCTs used**										

1.1. Goal setting (behaviour), 1.2. Problem solving, 1.3. Goal setting (outcome), 1.4. Action planning, 1.5. Review behaviour goal(s), 1.6. Discrepancy between current behaviour and goal, 1.7. Review outcome goal(s), 1.8. Behavioural contract, 1.9. Commitment, 2.1. Monitoring of behaviour by others without feedback, 2.2. Feedback on behaviour, 2.3. Self-monitoring of behaviour, 2.4. Self-monitoring of outcome(s) of behaviour, 2.5. Monitoring of outcome(s) of behaviour without feedback, 2.6. Biofeedback 2.7. Feedback on outcome(s) of behaviour, 3.1. Social support (unspecified), 3.2. Social support (practical), 3.3. Social support (emotional), 4.1. Instruction on how to perform the behaviour, 4.2. Information about Antecedents, 4.3. Re-attribution 4.4. Behavioural experiments, 5.1. Information about health consequences, 5.2. Salience of consequences 5.3. Information about social and environmental consequences, 5.4. Monitoring of emotional consequences, 5.5. Anticipated regret, 5.6. Information about emotional consequences, 6.1. Demonstration of the behaviour, 6.2. Social comparison, 6.3. Information about others' approval, 7.1. Prompts/cues, 7.2. Cue signalling reward, 7.3. Reduce prompts/cues, 7.4. Remove access to the reward, 7.5. Remove aversive stimulus, 7.6. Satiation, 7.7. Exposure, 7.8. Associative learning, 8.1. Behavioural practice/rehearsal, 8.2. Behaviour substitution, 8.3. Habit formation, 8.4. Habit reversal, 8.5. Overcorrection, 8.6. Generalisation of target behaviour, 8.7. Graded tasks, 9.1. Credible source, 9.2. Pros and cons, 9.3. Comparative imagining of future outcomes, 10.1. Material incentive (behaviour), 10.2. Material reward (behaviour), 10.3. Non-specific reward, 10.4. Social reward, 10.5. Social incentive, 10.6. Non-specific incentive, 10.7. Self-incentive, 10.8. Incentive (outcome), 10.9. Self-reward, 10.10. Reward (outcome), 10.11. Future punishment, 11.1. Pharmacological support, 11.2. Reduce negative emotions, 11.3. Conserving mental resources, 11.4. Paradoxical instructions, 12.1. Restructuring the physical environment, 12.2. Restructuring the social environment, 12.3. Avoidance/reducing exposure to cues for the behaviour, 12.4. Distraction, 12.5. Adding objects to the environment, 12.6. Body changes, 13.1. Identification of self as role model, 13.2. Framing/reframing, 13.3. Incompatible beliefs, 13.4. Valued self-identify, 13.5. Identity associated with changed behaviour, 14.1. Behaviour cost, 14.2. Punishment, 14.3. Remove reward, 14.4. Reward approximation, 14.5. Rewarding completion, 14.6. Situation-specific reward, 14.7. Reward incompatible behaviour, 14.8. Reward alternative behaviour, 14.9. Reduce reward frequency, 14.10. Remove punishment, 15.1. Verbal persuasion about capability, 15.2. Mental rehearsal of successful performance, 15.3. Focus on past success, 15.4. Self-talk, 16.1. Imaginary punishment, 16.2. Imaginary reward, 16.3. Vicarious consequences. PA = Physical activity targeted by app, PA+ = Physical activity and other behaviours targeted by app.

*Not downloaded. Additional evidence from academic papers suggested that additionally, 1.2, 1.4, 1.5, 1.6, 1.9, 3.1, 4.1, 4.2, 5.1, 5.3, 5.4, 5.6, 6.2, 8.2, 9.1, 15.3 also featured within the app combined with the Jawbone wearable. As the wearable was excluded from the present app evaluation and these additional BCTs were not found in the primary paper, they were not included.

$ Additional evidence from academic papers suggested that additionally, 1.5, 1.7, 2.6, 3.3, 5.4, 15.3 also featured within the app combined with the Fitbit wearable. As the wearable was excluded from the present app evaluation and these additional BCTs were not found, they were not included.

Greyed out cells indicate that an additional linked resource provides these BCTs, not the app itself, e.g. website.

Yellow cells indicate BCTs were focused on another behaviour, not PA. As they are included within an app that promotes PA, and mechanism of effect was not specified, these were considered relevant for inclusion.

Where 12.5 has been reported, this is addition of objects beyond the app e.g. wearable or scales (Health Mashups)

Green, red and blue main BCT labels indicate those with evidence for improving, reducing PA levels, or both (mixed evidence), respectively.

"+" depicts BCT is present in all probability, but evidence unclear. "++" depicts BCT is present beyond all reasonable doubt.

5.4.9.2 Physical activity recommendations and other evidence-based measure outcomes

Nine apps included or used guidelines/recommendations for activity levels, and four of those were public apps. Five literature-based apps used recommendations to make suggestions for user goals or provide feedback on user activity levels (MAPS, Motimate), the latter citing the Australian National Physical Activity Guidelines for Adults. One public app stated in its training sessions that users would have their activity levels brought up to the *'recommended levels'* and explicitly reported the UK's national recommendations for adults on its website (One You). The other public apps referred to the American Heart Association recommendations when users first set a goal (Fitbit), the Centre for Disease Control recommended activity levels on its webpage (Fitbit), adult and young adult recommendations for activity (Clue) and the World Health Organisation activity recommendations (Samsung Health).

No further feedback on evidence-based content was provided for literature-based apps, but five public apps reported the number of BCTs identified from other academic assessments of them. In each case, these additional assessments identified fewer BCTs than those identified in the present study (see table 27), ranging from a difference of 5 to 18 BCTs (see table 28 for further details).

Table 28 PA recommendations and evidence-based content in apps

<table>
<tr><th colspan="2">App</th><th>Presence of PA recommendations for adults (y/n)</th><th>Additional evidence</th></tr>
<tr><th colspan="4">Literature-based apps</th></tr>
<tr><td colspan="2">Haptivity (Not available for download) PA</td><td>No</td><td>NA</td></tr>
<tr><td rowspan="4">MAPS (Not available for download) PA</td><td>1</td><td rowspan="4">Used to make goal suggestions for aerobic and non-aerobic activity. Goal setting sliders are formed in days and minutes which match with format of recommendations in app conditions 2 and 4 (see table 17). Recommendations also cited during orientation. Goal setting urged progression toward public health recommendations for PA over the course of the 12-week intervention.</td><td rowspan="4">NA</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td colspan="2">Unnamed app (Not available for download) PA+</td><td>No</td><td>NA</td></tr>
<tr><td colspan="2">ATHENA apps (Not available for download) PA+</td><td>Not reported</td><td>NA</td></tr>
<tr><td colspan="2">Health Mashups (Not available for download) PA+</td><td>No</td><td>NA</td></tr>
<tr><td colspan="2">iN Touch</td><td>Not reported</td><td>NA</td></tr>
</table>

(Not available for download) PA+		
Motimate (Not available for download) PA+	Australian National Physical activity guidelines for adults used to provide text feedback, but unclear if guidelines are reported in app.	NA
Ngala (Not available for download) PA+	Not reported	NA
SIGMA (Not available for download) PA+	Not reported	NA
Descriptives:	**5/12 apps included national recommendations for PA**	**No feedback on evidence-base provided**
Public apps		
One You (Downloaded) PA	Within the app, during a run session, the trainer states that they will bring users up to running 30 minutes a day, 3 times a week. This is the recommended amount of MVPA, but it's not reported as such. The website reports 150 minutes MVPA or 75 minutes VPA a week recommended for adults with/without disabilities and describes what moderate PA is. Suggests breaking up sedentary time and building strength and balance.	Reports app includes 12 BCTs and no studies of effectiveness are associated with the app (Bondaronek et al., 2018)
Runkeeper	No	Reports app includes 12 BCTs (Direito et al., 2014)

(Downloaded) PA		Reports app includes 8 BCTs (Middelweerd et al., 2014)
Runtastic HR (Downloaded) PA+	No	NA
Runtastic (Downloaded) PA+	No	Reports app includes 8 BCTs (Bondaronek et al., 2018)
Runtastic PRO (Downloaded) PA+	No	Reports app includes 9 BCTs (Bondaronek et al., 2018)
Fitbit (Downloaded) PA+	American Heart Association recommendations are mentioned in app when setting an exercise goal, but cannot be accessed again after goal is set. Centre for Disease Control recommendations of 150 minutes of moderate exercise per week mentioned on help page on website (linked from app). Also a blog entry is available on the change to the guidelines.	Reports app includes following 16 BCTs (based on screenshots only): 1.1 Goal setting, 1.3 Goal setting (outcomes), 1.5 Review behaviour goals, 1.7 Review outcome goals, 2.2 Feedback on behaviour, 2.6 Biofeedback, 2.7 Feedback on outcomes of behaviour, 3.1 Social support (unspecified), 3.3 Social support (emotional), 5.4 Monitoring of emotional consequences, 6.2 Social comparison, 8.7 Graded tasks, 10.3 Non-specific reward, 10.4 Social reward, 10.10 Reward (outcome), 15.3 Focus on past success (Lyons et al., 2014). Reports app includes 17 BCTs but 40 BCTs used by users (not necessarily app based). BCT use not associated with PA level change. Also very little evidence for dose-response relationship. There was no evidence to support the hypothesis that the use of BCTs that are not explicitly included in the design of the Fitbit system would

		be positively associated with change in physical activity (daily steps and daily minutes of MVPA) over time. Participants who used the Social Comparison technique had a predicted increase in activity outcome over time - but most users used this. However, users were already very active (Ramirez et al., 2016). Reports app includes 10 BCTs (Bondaronek et al., 2018). Reports app includes 4 BCTs (Middelweerd et al., 2014). Reported app as meeting 15% of criteria for evidence-based behavioural weight loss strategies and 14% of technology assisted strategies (Pagoto et al., 2013).
Keep (Downloaded) PA+	No	NA
Clue (Downloaded) PA+	Under the information item for entering exercise data it reports the adult and adolescent recommendations: *"Healthy adults aged 18-65 years need moderate-intensity aerobic PA for a minimum of 30 minutes for 5 days a week or VPA for a minimum of 20 minutes for 3 days each week. Healthy adolescents aged 14-17 need at least 60mins of MVPA on most days."*	NA
Samsung Health (Downloaded) PA+	Under HR monitoring information tab it provides WHO recommendations for aerobic exercise for adults.	NA
Up	Not reported	NA

(Available version not functioning) PA+		
Descriptives	**4/10 apps included national recommendations for PA**	**5/10 apps reported feedback on evidence-based content**
TOTAL	**9/22 apps included national recommendations for PA**	**5/22 apps reported feedback on evidence-based content**

BCT = Behaviour Change Technique, HR = Heart Rate, MVPA = Moderate to Vigorous Physical Activity, NA = Not applicable, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app, VPA = Vigorous Physical Activity, WHO = World Health Organisation

5.4.10 Functionality/usability

Quality indicator	Summary definition
Functionality/Usability	Ease of use of the app and/or smartphone features, such as navigation, terminology, design in relation to ease of use, (not aesthetics) as well as general perception of how much support might be required for use, or how complex or inconsistent it might be. Functional errors related to app operations such as bugs/crashing also captured here. Includes practicality of use for promoting or capturing physical activity based on functions and features. Can be assessed by questionnaires such as the System Usability Scale, interviews or user-testing/performance tests.

Public and literature-based apps mean MARS scores were 4.39 and 3.95 respectively. However, there was a lot of missing data for the latter set. Apps that promoted multiple behaviours scored 4.25 compared to apps promoting PA alone (4.00). No individual app scored below average (see table 29 and appendix 18 for additional evidence that informed MARS scores).

Fourteen apps reported functionality outcomes, half of which were public apps. Apps were reported as easy to use, learn to use or required minimal effort (MAPS, Health Mashups, iN Touch, One You, Runtastic PRO, Fitbit, Clue), sometimes due to automated features being available (SIGMA, Fitbit). User-friendly designs were cited for two apps (Runkeeper, Up). Two apps were assessed for how many ergonomic criteria they met for each in-app task. Both Runkeeper and Runtastic met many of the criteria, but Runkeeper was considered easier to use for new users. The Up app was considered more time consuming for completing tasks than Fitbit, and more prone to prompting user error. Technical issues, or problems with remembering how content worked, were encountered for three apps (Health Mashups, Runtastic PRO, Up). Two of these aforementioned apps and others, were reported as unintuitive, difficult to use or understand, or content were easily forgotten/missed (Health Mashups, Ngala, SIGMA, Up) (see appendix 18).

Table 29 Functionality of apps

App name (downloaded or not)		Performance	Ease of use	Navigation	Gestural design	Mean score
Literature-based apps						
Haptivity (Not available for download) PA		NR	5	NR	NR	5.00
MAPS (Not available for download) PA	1	NR	3	NR	NR	3.00
	2	NR	4	NR	NR	4.00
	3	NR	3	NR	NR	3.00
	4	NR	4	NR	NR	4.00
Unnamed app (Not available for download) PA+		NR	3	NR	NR	3.00
ATHENA apps (Not available for download) PA+		NR	NR	NR	NR	NR
Health Mashups (Not available for download) PA+		NR	5	NR	NR	5.00
iN Touch (Not available for download) PA+		NR	4	NR	NR	4.00
Motimate		NR	4	NR	NR	4.00

(Not available for download) PA+					
Ngala (Not available for download) PA+	NR	NR	NR	NR	NR
SIGMA (Not available for download) PA+	NR	5	4	NR	4.50
Mean	**NA**	**4.00**	**4.00**	**NA**	**3.95**
Median	**NA**	**4**	**4**	**NA**	
Public apps					
One You (Downloaded) PA	4	5	5	5	4.75
Runkeeper (Downloaded) PA	5	3	4	5	4.25
Runtastic HR (Downloaded) PA+	4	5	4	5	4.50
Runtastic (Downloaded) PA+	5	4	4	5	4.50
Runtastic PRO (Downloaded) PA+	5	4	4	5	4.50
Fitbit (Downloaded) PA+	4	4	4	5	4.25

Keep (Downloaded) PA+	4	4	4	5	4.25
Clue (Downloaded) PA+	5	3	4	4	4.00
Samsung Health (Downloaded) PA+	4	5	4	5	4.50
Up (Available version not functioning) PA+	NR	NR	NR	NR	NA
Mean	**4.44**	**4.11**	**4.11**	**4.89**	**4.39**
TOTAL Mean	**4.44**	**4.05**	**4.10**	**4.89**	**4.16**
Median	**4**	**4**	**4**	**5**	
TOTAL Median	**4**	**4**	**4**	**5**	

NR = Not reported, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.11 Security and privacy

Quality indicator	Summary definition
Security/Privacy	Data privacy and/or security. For example, could include availability and accessibility of a privacy policy as well as its content, or a required login for the app or ability to make personal content private rather than sharing with app community. Could include meeting Data Protection standards.

Most of the apps explicitly reported a variety of security or privacy features (n=16). Volume and type of features employed varied between literature-based apps, with two apps reporting only the need to create an account or enter a postcode to sign in (Motimate, Ngala). Three others implemented multiple features including user registration, user permission to share data to social platforms, avoidance of analysis that compromised privacy, the ability for users to toggle on or off various tracking features, authentication requirements to link to app servers and other devices, data backup, use of unique usernames, and so on (ATHENA, Health Mashups, SIGMA). One app was also Health Insurance Portability and Accountability Act accredited (iN Touch). Public apps tended to include similar features to each other, with all except one explicitly providing a Privacy Policy (Up), and many facilitating user-managed data or profile visibility, marketing messages or data sharing (Runkeeper, Runtastic HR, Runtastic, Runtastic PRO, Fitbit, and Samsung Health). Most apps required a login. However, they remained logged in when users exited the apps. Few reported the ability to backup or download data (Runkeeper, Clue, and Samsung Health). One public app didn't explicitly report privacy/security features or have additional evidence that indicated they implemented any (Up).

Public apps reported additional evidence on security and privacy features. Runtastic, One You, Runtastic PRO and Fitbit were each scored on seven privacy criteria and scored 2, 4, 5 and 6 out of 7 respectively. Fitbit users reported liking the ability to compare data with anonymised users (see table 30 for further details).

Table 30 Security and privacy features in apps

App name (downloaded or not)		Features demonstrated	Additional evidence
Literature-based apps			
Haptivity (Not available for download) PA		User's posts were seen publicly. No indication of any other privacy/security features.	NA
MAPS (Not available for download) PA	1	None reported.	NA
	2		
	3		
	4		
Unnamed app (Not available for download) PA+		None reported.	NA
ATHENA apps (Not available for download) PA+		Users must register to use the platform. Sharing data to social networks was available with user permission. To maintain privacy users were assigned identities. (Fahim et al., 2014) Comparing all user activities to prescribed activity schedules and providing feedback via Focused Activities tab was considered computationally intensive, prone to compromising	NA

	the privacy of users, and not important from a healthcare point of view, therefore it was not performed. (Saleem et al., 2012) Trustworthiness is ensured by keeping the results of Social Media and Interaction Engine (SMIE) as unpublished knowledge in the social media repository, until it is approved by the clinicians before sending any recommendation to a patient. (Fatima et al., 2015) The architecture includes mobile services to support the secure transmission and processing of data in addition to the collection of other sensory data available from the mobile platform. A suite of mobile services will be developed to ensure the secure processing and transmission of all data collected from the users. These services will be responsible for managing security, efficient transmission of data and interfacing with cloud services. A data management module ensures the data is appropriately structured and formatted to ensure efficient transfer and storage. In this respect the security of the sensitive data is crucial, therefore efficient cryptography protocols shall be employed. It is envisioned that data could then be encrypted before being transmitted and stored in the cloud. (Cleland et al., 2013)	
Health Mashups (Not available for download) PA+	**Pilot outcomes** Background contextual data capture (e.g. location) could be turned off by users. Withings and Fitbit required user permission via OAuth to access their data for the MashUp. Therefore an account setup website was created specifically for this. OAuth allows private data sharing between services,	NA

	using a token rather than password system, for a specific data type (step counts and hours slept), for a specific duration. The Health Mashups Server is cloud computing. Data was backed up incrementally every 15 minutes in an Amazon server, and full backups executed nightly. Aggregate data from sensors, rather than raw data, is stored on the server as there is less risk if the server should be compromised, as only data summaries will be visible and not, for example, detailed time-stamped location data (Tollmar et al., 2012).	
iN Touch (Not available for download) PA+	TheCarrot, the app design platform, was HIPAA compliant (Health Insurance Portability and Accountability Act - for data privacy and security provisions for safeguarding medical info). Although all participants agreed their physician at the clinic and school nurse could access the weekly summary report (this was included in the consent form), a feature allowing patient's approval of provider access was also created as this was considered important for future implementations. iPod Touch required users to set a password for security. (Kim et al., 2015)	NA
Motimate (Not available for download) PA+	An account has to be created, but no further details required (Brindal et al., 2016).	NA
Ngala (Not available for download) PA+	Postcodes entered to sign in. (Hearn et al., 2014)	NA
SIGMA (Not available for download)	With respect to data protection, in addition to user authentication via username and password, all data protection issues will be covered by (1) having the locally stored data	NA

PA+	written in binary files that are difficult to alter, (2) by ensuring a secure HTTPS data transfer protocol, (3) by having a server authentication of the researchers, and (4) by using user aliases accessed by authorised personnel only. Furthermore, the customized feedback reports of each participant will only be available to themselves and protected by means of unique usernames and passwords (Podina et al., 2017).	
Descriptives:	**7/12 apps explicitly reported implementation of privacy/security features**	**0/12 apps received feedback on privacy/security features**
Public apps		
One You (Downloaded) PA	Privacy Policy available, refers to DPA and GDPR and details data processing. Terms and Conditions link. Both available via app. Joining the community requires personal information to be entered, but it can be hidden from the community and an anonymous username is advised.	App reported to meet 5/7 privacy criteria (has a Privacy Policy, which is available without having to download the app, as well as being available after the app is downloaded, collects personally identifiable information and shares data with 3rd parties) but did not have a short form Privacy Policy notice, and the policy wasn't available in other languages (Bondaronek et al., 2018).
Runkeeper (Downloaded) PA	Privacy Policy and Terms and Conditions available via app which links to website. Refers to a DP Officer. GDPR and Data Protection Act not mentioned. Potentially due to company being non-European (US, Japan, South Korea). Has ability to login/logout but remains logged in. User can manage how visible activities and maps are. Can export data from the website.	NA
Runtastic HR (Downloaded) PA+	Privacy Policy available on registration and includes reference to GDPR, DP officer, data processing, storage, user rights. Terms and Conditions and privacy policy also available in app	NA

<table>
<tr><td></td><td>under profile.
User can manage receipt of email and push messages on registration and marketing messages.
User can manage visibility of data, and their inclusion in leader boards.
Can create account to log in and out.</td><td></td></tr>
<tr><td>Runtastic
(Downloaded)
PA+</td><td rowspan="2">Privacy Policy available at website and on app, refers to GDPR and DP Officer, Data processing, sharing, storage, and user rights.

User can manage marketing messages.

User can manage how visible profile, activities and maps, photos and other components are.

User can manage their inclusion in leader boards.

Must register an account

Has ability to login/logout but remains logged in.

Runtastic stores data in a public or private cloud for data analysis as of 2013.(Ullrich et al., 2016)</td><td>App reported to meet 2/7 privacy criteria. It had a Privacy Policy, which is available without having to download the app but the policy wasn’t available after the app was downloaded. It didn’t collect personally identifiable information, did not have short form Privacy Policy notice, the policy wasn’t available in other languages and it didn’t share data with 3rd parties. (Bondaronek et al., 2018)</td></tr>
<tr><td>Runtastic PRO
(Downloaded)
PA+</td><td>App reported to meet 4/7 privacy criteria. It had a Privacy Policy, which is available without having to download the app. It collected personally identifiable information, and shares data with 3rd parties but the policy wasn’t available after the app was downloaded, did not have a short form Privacy Policy notice, and the policy wasn’t available in other languages. (Bondaronek et al., 2018)</td></tr>
<tr><td>Fitbit
(Downloaded)
PA+</td><td>Privacy Policy and Terms of Service available on the website and in app. Mentions GDPR and DP officer. Separate Privacy Policy for children’s data if parent creates account. Additional</td><td>The men liked comparing their physical activity anonymously with other participants via the pseudonym identifier on the</td></tr>
</table>

	details provided in app about privacy of female health tracking data. Must register to use app. Has ability to login/logout but remains logged in. User can manage access from third party apps and devices. User can manage visibility of data.	companion app. (Eisenhauer et al., 2017) App reported to meet 6/7 privacy criteria (has a Privacy Policy, which is available without having to download the app, as well as being available after the app is downloaded, collects personally identifiable information, shares data with 3rd parties and had a short form Privacy Policy notice) but the policy wasn't available in other languages (Bondaronek et al., 2018).
Keep (Downloaded) PA+	Privacy Policy and Terms and Conditions available on app, refers to GDPR. Privacy settings in app allows profile to be public/private. Must register to use app. Has ability to login/logout but remains logged in.	NA
Clue (Downloaded) PA+	Has Privacy Policy that also states passwords will be encrypted. Has a DP officer. Can use the app without an account, but data will only be on device. Has ability to login/logout but remains logged in. Can backup data. Website reports that data collected will be used for scientific research, but processed anonymously.	Includes password protection, data backup and has email or export data capacity. (Moglia et al., 2016)
Samsung Health (Downloaded) PA+	Privacy Policy available via link when first opening. Policy available online and refers to a European Data Protection officer as Data Protection laws change between countries where Samsung app is available. Can access Privacy Policy and Terms and Conditions via app under settings.	NA

	User can manage how data is used, such as for health data processing, shared with third-party apps. User can manage receipt of customised services and marketing information. Can set a password. Can download data to phone or erase data. Website suggests an award-winning security platform is used - Samsung Knox.	
Up (Available version not functioning) PA+	Not reported.	NA
Descriptives	**9/10 apps explicitly reported implementation of privacy/security features**	**5/10 apps reported feedback on security/privacy features**
TOTAL	**16/22 apps explicitly reported implementation of privacy/security features**	**5/22 apps reported feedback on security/privacy features**

DP = Data Protection, DPA = Data Protection Act, GDPR = General Data Protection Regulation, HTTPS = Hypertext Transper Protocol Secure, NA = Non applicable, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.12 Theoretical underpinnings and components

Quality indicator	Summary definition
Theoretical underpinning/components	Use of theoretical constructs from established behaviour change theories, or mention of use of theory, to inform development/content/evaluation of app in some way.

Literature-based apps reported more evidence of theoretical underpinnings than public apps. Five apps mentioned a theory, four of which were literature-based. Theories included Habit Theory, Self Determination Theory (reported by two apps), Cognitive Evaluation Theory, Social Cognitive Theory (reported by two apps), Self-Regulation Theory, the Health Action Process Approach, Conservation of Resources and the Cognitive Behavioural Theory including the ABC model. Four reported evidence that a theory or construct predicted behaviour, three of which were literature-based. Two literature-based apps were informed by more than a single theory, four of the same category of apps used theory to develop app intervention techniques and no apps used theory to select or tailor the app to recipients. Few apps reported app techniques being explicitly linked to a theory or the reverse (n=3 and n=2 respectively) (see table 31).

Table 31 Theoretical underpinnings of apps

<table>
<tr><th colspan="2">App</th><th>Theories and constructs</th><th>Techniques linked to theories/ constructs (All or at least one, How)</th><th>Theories/constructs linked to techniques (All or at least one)</th></tr>
<tr><td colspan="5">Literature-based apps</td></tr>
<tr><td colspan="2">Haptivity (Not available for download) PA</td><td>Mentioned: Yes.

Mentioned as predictor of behaviour: Don’t know.

Intervention based on <>1 theories: >1

To select recipients: No.

Develop intervention techniques: Don’t know.

Tailor techniques to recipients: No.</td><td>No.</td><td>No.</td></tr>
<tr><td rowspan="4">MAPS (Not available for download) PA</td><td>1</td><td rowspan="4">Mentioned: Yes.

Mentioned as predictor of behaviour: Yes.

Intervention based on <>1 theories: =1

To select recipients: No.

Develop intervention techniques: Yes.

Tailor techniques to recipients: No.</td><td rowspan="4">At least one – don’t know if grouped.</td><td rowspan="4">At least one.</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
</table>

Unnamed app (Not available for download) PA+	Mentioned: No. Mentioned as predictor of behaviour: No. Intervention based on <>1 theories: 0 To select recipients: No. Develop intervention techniques: No. Tailor techniques to recipients: No.	NA	NA
ATHENA apps (Not available for download) PA+	Mentioned: No. Mentioned as predictor of behaviour: No. Intervention based on <>1 theories: Don’t know. To select recipients: Don’t know. Develop intervention techniques: Yes. Tailor techniques to recipients: No.	NA	NA
Health Mashups (Not available for download) PA+	Mentioned: No. Mentioned as predictor of behaviour: No. Intervention based on <>1 theories: 0 To select recipients: No.	NA	NA

	Develop intervention techniques: No. Tailor techniques to recipients: No.		
iN Touch (Not available for download) PA+	Mentioned: No. Mentioned as predictor of behaviour: No. Intervention based on <>1 theories: 0 To select recipients: No. Develop intervention techniques: No. Tailor techniques to recipients: No.	NA	NA
Motimate (Not available for download) PA+	Mentioned: Yes. Mentioned as predictor of behaviour: Yes. Intervention based on <>1 theories: >1 To select recipients: Don’t know. Develop intervention techniques: Yes. Tailor techniques to recipients: No.	None linked.	None linked.
Ngala (Not available for	Mentioned: No. Mentioned as predictor of behaviour: No.	None linked.	NA

download) PA+	Intervention based on <>1 theories: 0 To select recipients: No. Develop intervention techniques: No. Tailor techniques to recipients: No.		
SIGMA (Not available for download) PA+	Mentioned: Yes. Mentioned as predictor of behaviour: Yes. Intervention based on <>1 theories: =1 To select recipients: No. Develop intervention techniques: Yes. Tailor techniques to recipients: Don't know.	At least one – not grouped.	All.
Descriptives:	**7/12 apps mentioned a theory** **6/12 apps mentioned a theory/construct as predictor of behaviour** **2/12 app were based on more than 1 theory** **0/12 apps used theory to select recipients** **7/12 apps used theory to develop intervention**	**5/12 apps explicitly linked behaviour change techniques to a theory**	**4/12 apps explicitly linked theories to behaviour change techniques used**

	Techniques **0/12 apps used theory to tailor the app to recipients**		
Public apps			
One You (Downloaded) PA	None reported.	None reported.	None reported.
Runkeeper (Downloaded) PA	None reported.	None reported.	None reported.
Runtastic HR (Downloaded) PA+	None reported.	None reported.	None reported.
Runtastic (Downloaded) PA+	None reported.	None reported.	None reported.
Runtastic PRO (Downloaded) PA+	None reported.	None reported.	None reported.
Fitbit (Downloade	None reported.	None reported.	None reported.

d) PA+			
Keep (Downloaded) PA+	None reported.	None reported.	None reported.
Clue (Downloaded) PA+	None reported.	None reported.	None reported.
Samsung Health (Downloaded) PA+	None reported.	None reported.	None reported.
Up (Available version not functioning) PA+	Mentioned: Yes. Mentioned as predictor of behaviour: Yes. Intervention based on <>1 theories: Don’t know. To select recipients: No. Develop intervention techniques: Don’t know. Tailor techniques to recipients: No.	At least one – group of techniques linked to group of constructs.	At least one.
Descriptives	**1/10 apps mentioned a theory** **1/10 apps mentioned a theory/construct as**	**1/10 apps explicitly linked behaviour change techniques to a theory**	**1/10 apps explicitly linked theories to behaviour change**

	predictor of behaviour **0/10 app were based on more than 1 theory** **0/10 apps used theory to select recipients** **0/10 apps used theory to develop intervention techniques** **0/10 apps used theory to tailor the app to recipients**		**techniques used**
TOTAL	**8/22 apps mentioned a theory** **7/22 apps mentioned a theory/construct as predictor of behaviour** **2/22 app were based on more than 1 theory** **0/22 apps used theory to select recipients** **7/22 apps used theory to develop intervention techniques** **0/22 apps used theory to tailor the app to recipients**	**6/22 apps explicitly linked behaviour change techniques to a theory**	**5/22 apps explicitly linked theories to behaviour change techniques used**

PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app

5.4.13 Usage and compliance

Quality indicator	Summary definition
Usage/Compliance	Not to be confused with engagement, usage or compliance refers to responses to app content, such a required step-count submissions, or response to prompts to complete questions. Similar to fidelity.

Thirteen apps reported usage data, nine of which were literature-based. Apps showed levels of usage in terms of the whole app, specific content or data entry, frequency of use over time or per day, population use, and gender differences in use, overall number of users and wear time of linked monitors. Apps tended to report positive usage patterns, where comparisons to other apps, version, controls, time points, or within users, was present (MAPS, Health Mashups, Motimate, Ngala, and Runtastic). Three apps reported generic usage – no comparison - (Haptivity, iN Touch, Fitbit), and one reported that PA and affect self-assessment and information resources were less used than those resources for other behaviours or feelings (Ngala) (see table 32 for further details).

Table 32 Usage and compliance with apps and their content

App		Usage results from primary paper and additional evidence*
Literature-based apps		
Haptivity (Not available for download) PA		Users were categorised into high or low frequency based on the number of photos they took, resulting in four high frequency users (9,10,21,22 photos), three low (3,4,7 photos). (Forster et al., 2017)
MAPS (Not available for download) PA	1	App use decreased across the intervention period (the random linear effect for time was significant (B =-.04, P<.01). The fixed effect for the goal setting module was also significant (B = .36, P = .02), as was the fixed effect for the points module (B = .38, P = .01), indicating the addition of either the goal or points modules was related to higher levels of usage (see table 17 for conditions). The interaction effects for time x points, time x goals, and points x goals were not significant, and were not retained in the model. Discussion refers to relatively high usage rates. (Fanning et al., 2017)
	2	
	3	
	4	
Unnamed app (Not available for download) PA+		NA
ATHENA apps (Not available for download) PA+		NA
Health Mashups (Not available for download) PA+		**Pilot outcomes** There was a significant difference in the amount of use of the website between the two participating countries. Sweden accessed an average of 70 pages while participants in Chicago accessed an average of only 10 pages (t (8) = 3.0, p < 0.03). No users visited web page that showed data sources for the day, where new data could be added/altered if there were issues. Across both cities users who walked more per day were more

	likely to use the mobile website. Overall, the use was less than expected, especially the frequency of manual logging. Many participants did not use the devices regularly enough to give examples of good, bad and average days, especially the manual logging. Across all participants, manual logging did not occur frequently. Food and Exercise were logged on average only 5 times by each participant over the 30 days of the second half of the study. Some data was only present for a few days, or on particular days of the week. Manual food logging was rarely used. In the first week, a few users tried it out, but after day seven, no more than two out of ten users logged food on the same day. After day 12, only one user sporadically logged food for the rest of the month. This resulted in an overall food logging rate of 12%. The rate of food logging behaviour per day increased by more than 5× in the full study (with reminders). The month of the pilot study, users averaged only a single day that had both Food and Weight logged. For the Steps and Food combination, users averaged 2.6 days in the pilot study. Sleep and pain were the least logged. Sensors weren't used as much as expected, users frequently reported not weighing themselves. Few users explored graphs from sensors. The automatic context streams of Calendar, Location, and Weather were quite bimodal, with many participants not using these sensors at all, and others who kept them on daily. **Trial outcomes** Higher usage than pilot, covering all aspects of the system and with no significant decline over time. Dramatic increase in logging reported. Most users that persisted beyond first 2 weeks used the system quite regularly for remaining 90 days. No usage differences between demographics (gender, location, education, and age) and usage was high. 63% of users logged food each day in the first month. This percentage stayed consistent throughout the month. Logging behaviour was sustained beyond the first month as well, with numbers between 50–70% each day during the second and third months of the trial. Pain was poorly reported across the board. The month of the pilot study, users averaged only a single day that had both Food and Weight logged, while in the full study participants averaged 9 such days in the first month and 21.3 over the full 90 days. For the Steps and Food combination, users averaged 14.8 days in the first month of the full study, and 37.9 days in the full three months. (Bentley et al., 2013; Tollmar et al., 2012)
iN Touch (Not available for	**Implementation outcomes** 2117 ODLs (observations of daily living) were recorded over 6 months. Across 24 users, 398 exercise, 1195

download) PA+	food (56.45%), 341 mood, 183 socialising ODLs captured. Per participant: range 1-699, mean 88.21. No. of days - range 1-179, mean 28.38. ODLs per day (including only days when at least one ODL was recorded) range 1-4.5, mean 3.11. Non-completing users recorded fewer ODLs (range 2-103, Mean 40.) (Kim et al., 2015)
Motimate (Not available for download) PA+	Negative binomial models indicated that those receiving Motimate remained active users of the app for 46 days longer than controls (P=0.017) (Brindal et al., 2018)
Ngala (Not available for download) PA+	Resource usage data obtained from Google Analytics (http://www.google.com .au/, verified 16 July 2013) showed 14,023 views of Ngala antenatal website pages and 7,596 of postnatal pages over the first year. A total of 2,378 users signed up to the Ngala app over the same time period. With an estimated 33,000 Western Australia (WA) births in 2012-13 and 40% of these being first time mothers, this extrapolated to 7% of pregnant WA women and 18% of first time mothers using the app. Website views and app new user sign up rates increased from the first six-month period to the second six-month period, with website views increasing to 72% and 14% in the antenatal and postnatal periods, respectively, and app user growth of 47% (62 new users per week increasing to 91 new users per week). Traffic to the Ngala website increased significantly after the launch of the Ngala web pages and app. Peaks in website traffic occurred directly after the launch and newspaper media coverage in July 2012, while a large increase in new app users was seen after newspaper publicity in May 2013. Direct traffic increased steadily since release of the Ngala resource. One quarter of website traffic was directly referred from the app. Ngala app self-assessments were completed at a rate of 167 per week, with the average person completing 3.6 questionnaires. The highest Ngala app self-assessment usage was in the first two trimesters of pregnancy and in the first 3 months after birth. The most popular Ngala app self-assessments completed over the year were weight (25%) and sleep (18%), followed by nutrition (15%), physical activity (15%), emotions (13%), and social life (13%). Aside from weight assessment, assessment topics were in similar proportions across pregnancy but with

	increasing emphasis on sleep until 9 months postpartum. The highest website page views related to the antenatal period featured nutrition content (40% of views) followed by weight (33%), physical activity (14%), sleep (6%), emotions (5%), and social life (2%). For page views related to the postnatal period and compared to antenatal, nutrition was again the highest focus (49% of views), with less views for weight (9%) and similar views for physical activity (16%), sleep (11%), emotions (11%), and social life (5%). When compared to the geographic distribution of mothers' residence for annual births WA regional use of the Ngala app was about 10% lower than in Perth but otherwise showed a similar distribution across regions. App users were resident in areas across all quintiles of social disadvantage but were over represented in areas of lower disadvantage compared to distribution of WA births. Whilst demographic data for Ngala webpage views was not available, registrations for the app showed slightly higher use in urban than rural areas. This was not unexpected since some WA rural areas have poor internet access or limited PHCPs to refer the app. (Hearn et al., 2014)
SIGMA (Not available for download) PA+	Assessed but outcomes not reported.
Descriptives:	**9/12 apps reported data on usage**
Public apps	
One You (Downloaded) PA	Usage not captured.
Runkeeper (Downloaded) PA	Usage not captured.
Runtastic HR (Downloaded) PA+	Usage not captured.
Runtastic	About one third of the sample (32.9%; n = 56) used the Runtastic app quite frequently, i.e. more than four

(Downloaded) PA+	times a week. Nearly half of the sample (48.8%; n = 83) used the app one to four times a week. The remaining 18.2% (n = 31) used Runtastic only once or twice a month. More than four fifths (85.8%) had used Runtastic for longer than 4 months, 2.9% had just started using the app within the last month, and 11.2% had used the app for between 2 and 4 months. There were no gender differences in the sample concerning duration of app usage. The three features of the app that were used most often to engage social support were sharing results, voice coach, and live tracking (Klenk et al., 2017).
Runtastic PRO (Downloaded) PA+	Usage not captured.
Fitbit (Downloaded) PA+	Participants reported interacting with app between 0 and 20 times per day (M = 2.94, SD = 3.81). For the website, participants reported logging on between 0 and 7 times per week (M = 1.75, SD = 2.33). (Ramirez et al., 2016)
Keep (Downloaded) PA+	Usage not captured.
Clue (Downloaded) PA+	Clue has more than 4 million users in over 180 countries tracking their cycles for a wide variety of reasons (as of October, 2015).
Samsung Health (Downloaded) PA+	Usage not captured.
Up (Available version not functioning) PA+	Although the intervention did not instruct usage of non-activity portions of the app, participants spontaneously tracked their sleep (Mean 11.70, SD 11.97 days) and food intake (Mean 2.65, SD 7.83 days). Of those who completed the study who were randomised to the intervention group, 13 out of the 21 participants (61.9%) reported interacting with the UpBand at least once per day. Average activity monitor wear time was 81.85 (SD 3.73) of 90 days with a minimum of 69 days. (Harris et al., 2018)
Descriptives	**4/10 apps reported data on usage**
TOTAL	**13/22 apps reported data on usage**

*Where possible, text is verbatim or reduced for length verbatim, or summarised using key terms from paper

NA = Non-applicable, ODL = Observations of Daily Living, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app, PHCP = Primary Health Care Providers, SD = Standard Deviation, WA = Western Australia

5.4.14 General quality

Public and literature-based apps had overall mean quality scores of 4.15 and 3.41 respectively, derived from MARS. Apps that promoted PA alone and apps promoting multiple behaviours received almost equivalent mean scores of 3.72 and 3.76 respectively. Free and paid public apps had means of 4.22 and 3.87 respectively. Only one app scored below average (a score of 3) on overall quality (Unnamed).

One literature-based app reported an additional quality measure: player experience, showing positive outcomes (SIGMA). However, given the recent conceptualisation of engagement, this could be incorporated under this criterion instead (Perski et al., 2017) (see table 33 for details).

Table 33 Overall quality of apps

App		Other quality measure results	MARS mean score
Literature-based apps			
Haptivity (Not available for download) PA		NA	3.78
MAPS (Not available for download) PA	1	NA	3.07
	2		3.65
	3		3.39
	4		3.88
Unnamed app (Not available for download) PA+		NA	2.35
ATHENA apps (Not available for download) PA+		NA	3.04
Health Mashups (Not available for download) PA+		NA	3.80
iN Touch (Not available for		NA	3.15

download) PA+		
Motimate (Not available for download) PA+	NA	3.68
Ngala (Not available for download) PA+	NA	3.25
SIGMA (Not available for download) PA+	Regarding player experience, analysis revealed a mean score of 74.23 (SD = 8.50; range 16-96), based on participants' PES-16 ratings. The minimum value was 60 (1 participant), while the maximum value was 89 (1 participant). All participants had scores above 60 on the PES-16. The analysis of the PES-16 ratings revealed that the participants (1) viewed the game experience as a reward of its own (M = 4.10, SD = 0.66); (2) perceived they had freedom to perform the desired action within the game (M = 4.13, SD = 0.91), and experienced increased focus and engrossed attention during the game (M = 4.83, SD = 0.84). While for the three dimensions of the PES-16 responses were grouped around a clear cut favourable perception of the game opinions were somewhat more toned down with regard to the utility of the game in the real world setting (M = 3.42, SD = 0.55). (Podina et al., 2018)	3.96
Descriptives:	**1/12 apps assessed an additional measure of quality**	**3.41**
Public apps		
One You (Downloaded) PA	NA	4.05
Runkeeper (Downloaded)	NA	4.30

PA		
Runtastic HR (Downloaded) PA+	NA	3.67
Runtastic (Downloaded) PA+	NA	3.98
Runtastic PRO (Downloaded) PA+	NA	4.08
Fitbit (Downloaded) PA+	NA	4.43
Keep (Downloaded) PA+	NA	4.40
Clue (Downloaded) PA+	Tech support was available (Moglia et al., 2016)	3.94
Samsung Health (Downloaded) PA+	NA	4.53
Up (Available version not functioning) PA+	NA	4.11
Descriptives	**1/10 apps assessed an additional measure of quality**	**4.15**
TOTAL	**2/22 apps assessed an additional measure of quality**	**3.75**

M = Mean, NA = Not applicable, PA = Physical Activity/Physical Activity targeted by app, PA+ Physical Activity and other behaviours targeted by app, PES = Player Experience Scale, SD = Standard Deviation.

5.4.15 User reviews

The 50 most recent user reviews since the app was downloaded for evaluation were extracted for eight of the public apps from the app stores. The Up app did not have reviews extracted as it was discontinued, and the version was not reported. One app only had 45 recent reviews at the time of extraction (Runtastic HR), resulting in there being a total of 445 reviews. Nonsense (e.g. *"spell Justin")* or non-English reviews were excluded and replaced with the next sequential review to ensure 50 reviews per app where possible.

Reviews were coded as either positive or negative. Ninety-seven were negative (mean 10.77, 21.7%), 277 positive (mean 31.33, 63.4%) and 71 were both (mean 7.89, 16%). See appendix 21 for number and type of reviews per app.

One hundred and thirty-seven simple, generic reviews that were either positive or negative were separated from more informative reviews (e.g. *"Great!"* or "*very bad*"). These were considered indicators of app acceptability, however. Eight themes were identified from the remaining 308 informative user reviews.

Seven themes mapped on to the previously defined quality criteria: Acceptability, Aesthetics, Credibility, Effectiveness/Potential impact, Engagement, Functionality/Usability and Privacy/Security. However subthemes included as yet unmentioned facets of these quality criteria. Not every theme was represented for each app, for example, Aesthetics was only mentioned, and praised, for three (Clue, Runtastic HR and Samsung Health). Security was only mentioned, and as a negative, for two (Runtastic PRO and Runtastic HR). A summary of the remaining themes is available in table 34. Subthemes and exemplar quotes are reported in appendix 22.

Most themes had corresponding positive and negative aspects, but not all of them, reflected where possible in subtheme names.

Table 34 Summary of themes from user reviews

Theme	Summary
Acceptability	All apps received reviews on this theme, both negative and positive. Positive reviews included compatibility of the app with devices (Runtastic HR), it being cost efficient/free (Keep, Runkeeper, Runtastic, Runtastic HR), that users would recommend it (Clue, Keep, One You, Runkeeper, Runtastic HR), that users liked specific content (all), or that the user had a positive experience with the app (all except Runtastic and Runtastic HR). Conflicting negative reviews for some apps occurred. They referred to hidden costs (Keep, Runkeeper, Runtastic, Runtastic PRO), the app being incompatible with devices (Clue, Fitbit, Runtastic HR, Runtastic PRO, Samsung Health), that users wouldn't recommend the app (Clue), that users wanted to remove or request certain features (all except Fitbit and Runtastic) and that the app was resource heavy (Samsung Health).
Credibility	All apps received reviews on this theme. However one received only positive reviews (Runtastic HR). Positive reviews cited app accuracy/reliability (all except Fitbit, Keep and One You), perceived trustworthiness or safety of the app (Clue, One You, Runtastic HR) and usefulness of the app (all). Conflicting negative reviews for some apps occurred, perceiving apps as inaccurate, including all of which had also been perceived as accurate (Clue, Runkeeper, Runtastic HR, Runtastic PRO, Samsung Health), unsafe (Clue, Keep) and unhelpful (Keep, One You).
Effectiveness	All except one app received positive reviews on their effectiveness (Runtastic HR). No negative reviews were made. Users perceived all remaining apps as motivational (except Clue) and noticed a physical effect, or change, in their behaviour that they associated with the app.
Engagement	Six apps received positive reviews on engagement in general (Clue, Keep, One You, Runkeeper, Runtastic, Samsung Health), with three receiving praise for their tailoring or suitability for users (Clue, Keep, One You).
Functionality/ Usability	Seven apps had mixed reviews regarding functionality. However, Fitbit and Runtastic only received negative reviews. Positive reviews for the remaining apps related to their ease of use, and an additional two apps were cited as working as they should (Runkeeper and Runtastic HR). Negative reviews predominantly focused on functional errors, but varied in terms of the number of apps that reported specific errors. Crashing was reported for One

	You, Runkeeper and Runtastic PRO. Only Fitbit was reported as having network errors. Two apps had problems with their syncing feature (Fitbit, Runkeeper). Issues with, or following, app updates were also reported widely (Clue, Fitbit, Runkeeper, Runtastic HR, Runtastic PRO, Samsung Health). All except Runkeeper and Runtastic were reported as being difficult to use and two apps were additionally reported as being misleading in what they provided (Runtastic, Samsung Health). *(It should be noted that functional issues were rarely, if ever, experienced by the student during testing.)*
Other	Six apps received a small number of reviews that didn't fit into any existing themes and were only prevalent enough to generate a single additional subtheme. Three apps reported problems with app customer support features (Keep, Runtastic HR and Runtastic PRO), including being unable to find contact details, or receiving no response.

5.5 Summary and synthesis of results: Narrative mixed research synthesis using segregated design

Synthesising these results required a return to the thesis research questions.

Research questions:

1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps that include feedback on immediate affect, including both apps developed for/by researchers and publicly available apps (commercial apps) in the app stores?
3. What is the quality of these apps that provide feedback on immediate affect, where quality is defined as a multi-faceted concept consisting of 13 features?

Findings from all the sources previously reported were assessed for whether they confirmed or refuted (where findings addressed the same facet of a quality criterion) or complemented (where findings addressed different facets of a single quality criterion) each other (Sandelowski et al., 2006).

5.5.1 Synthesis for RQ1: Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?

Twenty-two physical activity apps were identified that included feedback on affect, however some were more explicit about using it to promote PA than others, based on descriptions of how affect was captured, processed and fed back (see 5.5.2).

5.5.2 Synthesis for RQ2: What are the characteristics and content of physical activity apps that include feedback on immediate affect, including both apps developed for/by researchers and publicly available apps (commerical apps) in the app stores?

Descriptive analysis of characteristics of the apps reported that no literature-based apps were developed outside of developed countries, or were available for download and full evaluation. A variety of methods were used to design and test them and few were evaluated using randomised designs. Some form of quality assessment was completed for most apps, but few issues were actively addressed. All public apps except one had been independently evaluated for effectiveness or quality.

Four apps promoted PA alone, while the remaining eighteen addressed multiple behaviours. Literature-based apps failed to report size and cost, but most of the included public apps were free. Complexity and volume of content varied extensively between apps and most allowed user-led or app-based tailoring. Public apps included more technical aspects than literature-based apps. Few explicitly reported targeting inactive individuals or those from BAME groups.

Six apps captured and gave feedback on PA-contingent affect, while the rest captured and fed back affect in the context of an app that happened to promote PA, as well as potentially promoting other behaviours. Self-report measures of affect were used predominantly, but different methods were employed, only two of which were reported as validated measures. Over half the apps made affect data more salient, with less preferring to repeat the data as entered when providing feedback. Only literature-based apps performed additional processing to increase the salience of affect data. Feedback on affect tended to be available for the duration the app was used.

5.5.3 Synthesis for RQ3: What is the quality of these apps that provide feedback on immediate affect?

5.5.3.1 Acceptability

While literature-based apps did not report user ratings or reviews, public apps did both. User ratings were uniformly above average, if not high. However, user reviews refuted the high scores for some apps, suggesting most were not completely acceptable and struggled with compatibility issues, the need for additional content, or the removal of content and excessive resource use. Similarly, evidence of acceptability from the literature reported previously lower user ratings, but only for one of the apps that received negative user reviews (Fitbit). Additional evidence complemented user ratings, showing mixed evidence for the acceptability of public apps, but although the same issues around cost and content appeared, they were only for a subset of the public apps receiving negative reviews.

User reviews confirmed evidence from the papers that suggested users liked specific app content and would recommend certain apps, although the public apps themselves didn't always match between the two sources. User reviews introduced additional facets of acceptability that were not captured in the papers, including compatibility with devices and apps being cost efficient.

5.5.3.2 Aesthetics

Public apps scoring highly for aesthetics, as derived by MARS, were complemented by the positive user reviews they received, although very few reviews referred to aesthetics. However, two of the three that received user reviews on this criterion achieved the second highest mean MARS score (Runtastic HR PRO, Samsung Health), with only one app scoring higher (Fitbit).

Literature-based apps varied in their scores for this criterion ranging from 2 to 4.33, suggesting this was an area where they struggled in comparison to the public apps (ranging from 4.33 to 5).

5.5.3.3 Physical activity measurement tool

Almost all apps captured or measured PA and out of the three apps that didn't, users requested it for one (Haptivity). Both self-report and objective measures were used, but public apps tended to use both, whereas literature-based apps used the former more frequently. While both types of measurement have established strengths and weaknesses, additional evidence often reported other strengths and weaknesses that the authors or users had identified. These complemented the established issues, reinforcing the mixed evidence for each measure in general, as well as providing very specific challenges and strengths for each individual tool. See 5.5.3.4 for further discussion of app accuracy.

5.5.3.4 Credibility

Mid-range MARS scores for this criterion were confirmed by mixed user reviews for public apps. User reviews and additional evidence confirmed that users appreciated the apps being useful and accurate or reliable. But user reviews also reported trustworthiness or safety of the apps as important, while additional evidence reported that apps informed learning.

User reviews confirmed the facets of credibility that were important to users and represented in the apps, including accuracy, safety and helpfulness. Their importance was further supported by additional evidence suggesting apps had mixed findings for accuracy and validity.

5.5.3.5 Currency and maintenance

While literature-based apps mostly failed to report release date and updates, public apps were up to date and demonstrated regular and ongoing maintenance since release for all except one which was discontinued (Up), and one that was updated less frequently (Keep).

5.5.3.6 Development process and team

Additional evidence and the availability of user reviews suggested that most apps involved potential users in the initial or ongoing design and development of the app. However literature-based apps suggested more significant involvement from initial design through to pilot testing. Most apps had commercial affiliations. Public and literature-based apps differed in the composition of their development teams with respect to technology and behaviour change or health experts. Additional evidence confirmed the lack of health experts and lack of user involvement in the initial development of public apps.

5.5.3.7 Effectiveness/Potential impact

While additional evidence and user reviews suggested apps were perceived as, or were effective for, either changing behaviour or motivating users, MARS scores somewhat refuted this by being mid-range to below average for at least half the apps.

5.5.3.8 Engagement

Overall, apps scored just above average for engagement (MARS), while limited evidence for engagement came from the literature or from user reviews. This suggested that apps rarely delivered salient, engaging, content that warranted extensive feedback from users.

5.5.3.9 Evidence-based content and components

Data generated for this thesis was refuted by evidence from the papers with respect to the number of BCTs reported in certain apps, with more BCTs being reported for the present study. Public apps tended to include more BCTs associated with improving PA levels than literature-based apps, but most apps included at least one of three that were associated with mixed evidence, namely 2.2 Feedback on behaviour.

Despite the extensive use of evidence-based content in the form of BCTs, presence of recommendations for PA were rarely used or referred to across the apps (n=6).

5.5.3.10 Functionality

Apps scored highly with respect to their functionality, where data was available to code MARS. However, user reviews and additional evidence suggests mixed experiences for users. More functional errors were reported for public apps in user reviews, but both reviews and additional evidence broadly reported that the set of apps in general were easy to use or user-friendly. Additional evidence reported that some apps were difficult to use, but different public apps were reported as such by user reviews.

5.5.3.11 Security and privacy

Public apps reported similar features within their group, but some different features to literature-based apps. Evidence for issues with security features was mixed between additional evidence and user reviews of public apps, but in both cases it was infrequently assessed.

5.5.3.12 Theoretical underpinnings

Limited evidence of extensive theoretical underpinnings were found across the apps, though literature-based apps exhibited them more frequently than public apps, when reported at all. There was little consistency with respect to the theories informing the apps and rarely were explicit links between theories and app content or techniques made.

5.5.3.13 Usage and compliance

Different methods of capturing usage were employed across the apps that reported it. Usage was typically good when comparisons were made.

However, where generic usage was assessed in one app, it suggested poor use of particular components compared to other components (Ngala).

5.5.3.14 General and other measures of quality

Public apps scored higher than literature-based apps with respect to overall quality (MARS). Apps promoting PA only and those promoting multiple behaviours faired equally.

Only one app was evaluated on an additional quality criterion not included in this thesis – player experience. However, user reviews confirmed that other quality issues may be pertinent to users, such as the availability of customer support, which was problematic for three public apps (Keep, Runtastic HR, Runtastic PRO).

5.6 Risk of bias

Threats to reliability and validity of data will be reported in Chapter 6, Discussion.

6. Discussion

6.1 Summary of evidence

Following a systematic literature review and a systematic identification of public apps from the largest two app stores, 22 apps were located that promoted physical activity and provided feedback on user affect. However, different apps were more explicit than others about both promoting physical activity and using feedback on affect as a potential facilitator for behaviour change. For example, some apps used multiple or overt strategies to promote physical activity, whereas others focused on multiple behaviours or symptoms, including physical activity. Many apps also provided feedback on affect in a generalised manner, not actively linked to physical activity, but in the context of an app that also promoted physical activity. Despite this, more than half of the apps processed their affect data to make it more salient to users when it was fed back, but literature-based apps were more inventive and arguably more informative. This suggested that their development teams were more aware of the importance of affect for user behaviour.

Apps were created in developed countries and only public apps were available for download, of which most were free. Few targeted inactive individuals or those from Black, Asian and Minority Ethnic (BAME) groups. Most studies describing apps reported their development and pilot testing, using a variety of methods. Therefore, some facet of quality was frequently assessed, but few reports demonstrated that changes were made based on these assessments.

The complexity of apps varied within both literature-based and public apps and this did not seem to depend on whether the app promoted physical activity alone or multiple behaviours. Multiple strategies were employed across apps in all cases. These strategies were reflected in their descriptions and their use of BCTs. Most apps incorporated both user-led (allowing user to select activities, goals etc.) and app-based tailoring (goals or activities automatically recommended, based on user entered data and characteristics). Public apps included more technical aspects than literature-based apps. Some of the aspects the literature-based apps neglected to include were those that potentially required higher level programming, or access to other services or resources such as the ability to share information on social media and the presence of an app community. Others, it could be argued, should have been as standard and, therefore, perhaps were not reported (password protection, login requirements).

The availability of data for discerning the quality of the included apps varied. As the individual and synthesised findings demonstrated, quality of the apps was mixed, either between or within the 13 criteria (acceptability, aesthetics, physical activity measurement, credibility, currency, development teams, effectiveness, engagement, evidence based content, functionality, security, theoretical underpinnings and usage). This is discussed in the subsequent sections.

6.2 Relationship to other evidence

6.2.1 Study designs, target groups and quality assessment methods

Given the purpose of the studies that reported literature-based apps, few randomised designs were used, as these are more appropriate for testing effectiveness, after sufficient levels of quality have been established. This fits with recent recommendations from a consort of experts, who suggested that RCTs were challenging for digital interventions and more flexible, quicker methods could be used (Michie et al., 2017). Factorial studies to assess individual digital intervention components were also advocated (Murray et al., 2016). However, only one app study used this design (Fanning et al., 2017) while neither of the proposals for the evaluation of two other apps (SIGMA, Motimate), incorporated this design. Instead, both reported a randomised trial design. It was also unfortunate that few literature-based apps appeared to have subsequently been rigorously tested for effectiveness. This may have been due to the recent publication of the included papers, or the lack of funding and support for further evaluation.

Few apps reported methods for usability testing that are preferred by user-experience experts. Many included apps were assessed using focus groups, interviews and questionnaires, as well as real-world testing by users outside a lab. Jakob Nielsen, a usability expert, reports for example, that focus groups are "*a poor way of evaluating design usability*" (Nielsen, 2012). This is because when dealing with interactive content, such as that from an app, it is important to observe what the user does. What they tell you they do, may be inaccurate or misleading. While additional evidence from evaluations of public apps did suggest that more traditional usability test methods were used, including those where apps were tested by users performing representative tasks (Altenhoff et al., 2015; Klock and Gasparini, 2015) and the System Usability Scale (Brooke, 1996), these were performed after the app was public.

There are observed differences in physical activity levels based on population characteristics such as age, race, gender and socio-economic status (Cooper et al., 2015; Hallal et al., 2012; Trost et al., 2002; Wardle and Steptoe, 2003; Williams et al., 2011), showing that older adult women of lower-socioeconomic status and from BAME groups should be targeted for intervention in order to reduce health disparities. Health app use is also most prevalent in younger, more educated users and in men, and younger users are more likely to use fitness apps in particular (Bol et al., 2018). Evidence from the present studies showed that most at-risk individuals were rarely targeted by apps that include feedback on affect, suggesting a gap in app provision. In addition, one study suggests that public apps target already active people, for example Runkeeper was highlighted as such an app (Bielik et al., 2012). While One You may appeal directly to those who have little experience of being active, apps like Keep, Runkeeper and Runtastic seem more accessible to those who have already established a training regime. Therefore, it would appear that most apps supported this claim, by inadvertently or not, targeting those who were already categorised as 'early action', 'maintainers' or even 'habituated', to regular exercise (Norman and Velicer, 2003). The lack of targeting in public apps reinforced the hypothesis that developers may be interested in financial gain, rather than behaviour change for those most at risk, or at least unaware of such at risk groups.

App development has predominantly taken place in developed countries. While this is a typical finding of reviews of certain types of health interventions, as well as smartphone based interventions (Bort-Roig et al., 2014; Zhao et al., 2016), it may be appropriate, as evidence suggests that the western and developed world have higher levels of inactivity than other parts of the world (Hallal et al., 2012).

The current evidence reinforces the claim that there is, as yet, not an agreed gold standard for evaluating behaviour change apps (BinDhim et al., 2014; Jeon et al., 2014; Michie et al., 2017).The quality criteria assessed and reported across the apps varied. Each app had some evidence for at least one of this theses criterion, but not all apps had data available for all criteria. For example, very few apps provided substantial, or any, data on theoretical underpinnings – this was the most infrequently reported criterion.

6.2.2 Feedback on affect and its use for promoting physical activity

The present findings support the claim made in the introduction of this thesis that there are few PA promotion apps that explicitly use feedback on affect as a technique to facilitate PA behaviour change. More specifically, there are fewer apps providing feedback explicitly on PA-contingent affect. The present findings suggest that the recognition of the importance of affect in behaviour change has not been fully translated to the PA app domain, despite the call for raising awareness of the link between affect and activity (Carels et al., 2007) and its use in apps (Stevens and Bryan, 2012). The effectiveness of these apps for raising awareness of the link between activity and affect has not been assessed, and the prevalence of multi-behaviour, multi-technique apps means that it is still difficult to confirm mechanisms of effect and theoretical basis for the relationship between affect and PA.

While some researchers may prefer objective measures of affective outcomes (and indeed many are becoming available, although their use may not be feasible for a free-living context (e.g. Neale et al., 2017)), the prevalent use of self-report measures of affect allowed for feasible data capture for users. As discussed in the introduction, self-monitoring has a strong evidence base for behaviour change and self-regulation, and is also evidenced to increase self-awareness of behaviour (Bandura, 1991; Dombrowski et al., 2012; Greaves et al., 2011; Harkin et al., 2016; Runyan et al., 2013; Tomita et al., 2008). Therefore, it could be argued that the included apps have been using the appropriate method of data capture to promote an increase in awareness and behaviour change. However, limited efforts have been made to make the relationship between physical activity and affect explicit to users, an area for improvement.

6.2.3 Quality of included apps

An overview of the findings from this mixed methods approach confirmed the existence of indicators of quality and their application to smartphone apps, for example evidence-based content. It also explored indicators derived from user feedback, for example acceptability. Due to the volume of data collected, it is challenging to discuss individual findings in relation to existing literature; but key findings for each criterion will be addressed. In addition, initial findings suggested that some criteria may be related, or that findings converged, reinforcing the credibility of the data, therefore discussions will be combined where appropriate.

Acceptability

While acceptability of the apps was mixed across the data sources, with a lack of corroboration between the different data sources at times, findings supported the acceptability and preferences of users for certain components, demonstrated by previous work. Tailoring, use of goals, monitoring, feedback, rewards, social support, as well as technical features specific to digital interventions, such as auto-pause settings, music integration, reminders, real-time logging, lack of advertisements and the ability to capture a range of behaviours have all been reported previously (Coughlin et al., 2016; Dennison et al., 2013; Middelweerd et al., 2015; Rabin and Bock, 2011; Raines, 2013; Vandelanotte and De Bourdeaudhuij, 2003). While extra evidence was lacking for its acceptability, auto-pause features have previously been perceived as beneficial to users (Voicu et al., 2019). Findings also supported evidence that users' desired recommendations and advice on how to improve behaviours, rather than merely being provided with statements on the performed behaviours, (Dennison et al., 2013) and for freely available apps (Dennison et al., 2013).

Aesthetics

Limited evidence was available for judging the quality of the apps in terms of their aesthetics. It was rarely reported outside of the data from the MARS score, and there was much missing data for the literature-based apps for this item. Unsurprisingly, public apps scored highly on this criterion due to their availability for evaluation, corroborating MARS scores and adding support to the existing evidence for the above-average aesthetic quality of public apps for behaviour change (Bardus et al., 2016; Schoeppe et al., 2017). In addition, literature-based apps did not include technology experts in development teams, or commercial or government affiliations to the same extent as public apps. The aesthetics of the app may have suffered due to these gaps in expertise and potential funding.

Credibility and the physical activity measurement tool

These criterion have been combined for discussion as user perspectives suggested a convergence on the importance of certain facets.

Both objective and self-report measures were used to capture various characteristics of physical activity (e.g. frequency, duration, intensity). While obvious strengths emerge for both self-report and objective measures, none of the self-report measures were validated and there was mixed evidence for the accuracy and reliability of objective measures, in particular inbuilt phone sensors. In addition, not all apps captured the intensity of PA, an important factor in PA for health with intensity specified in recommendations (Department of Health and Social Care, 2019; World Health Organisation, 2010). This could reduce the effectiveness of the app. Where intensity was captured by self-report, users also may have over or under-estimated, a common problem in PA research (Sallis and Saelens, 2000).

The current findings support the use of both self-report and objective measures of different characteristics of physical activity using smartphones, but also support the variation in measurement accuracy of objective sensors and devices (Bort-Roig et al., 2014; Case et al., 2015). Unfortunately, users frequently cited accuracy and reliability when leaving reviews. While evidence for accuracy was poor or mixed for a number of public apps, (corroborated by MARS scores for Credibility), users often cited accuracy and reliability when reviewing them, demonstrating the importance of this facet. This desire for accuracy both in terms of tracking and information provided, has been evidenced already in the app literature and poor reliability has shown to impact motivation, making it a problem for these apps (Baretta et al., 2018; Bickmore et al., 2009; Dennison et al., 2013).

Despite these issues, on the whole, apps were reported as having above average credibility scores. Many components that had been cited as being acceptable, both in the present study and in the literature (see above), were also perceived by users as being helpful, or useful, including recommendations, goals, feedback and in-app information. This builds on existing evidence, supporting ideas for future apps and their components (Bort-Roig et al., 2014).

Currency, maintenance and the development process

While evidence for the impact of recent updates on downloads is scarce (Nayebi et al., 2016), a Google search of recommendations for update frequencies for developers returns numerous web pages. They suggest that successful apps release updates up to four times a month, depending on user reviews and resources (Pisuwala, n.d.). Updates were also said to help to build a "*loyal following*" and demonstrate that the app has not been discontinued (Pisuwala, n.d.; Yarmosh, 2016). However, there's also the suggestion that updates can irritate users, as they consume costly data resources (Pisuwala, n.d.). As suspected, literature-based apps were typically not maintained, despite the publication of papers within the last 10 years. This is concerning in terms of wasted resources, as well as wasted, potentially effective and untested, interventions. Public apps, however, appeared to be routinely maintained with updates occurring less than a month before they were evaluated. As cost seems to be an acceptability issue for apps, regular updates at the rate demonstrated by most of the eligible apps may not be preferred by users. However, a standard smartphone feature now allows users to prevent automatic updates, reducing this risk.

Findings from the development process give insights into why there is a dichotomy between literature-based and public apps in terms of maintenance. Technology experts were infrequently cited as part of the development team for literature-based apps, whereas public apps were developed by companies whose business is app development. In contrast, literature-based apps included behaviour change, health or activity experts in their development teams and more frequently reported user-involvement too. The latter is likely due to the requirement for patient and public involvement (PPI) in health research from many large funding agencies (e.g. National Institute for Health Research, Medical Research Council), as well as the release of PPI national standards (National Institute for Health Research, 2017). These findings suggest that recommendations for PPI are being adhered to (Michie et al., 2017). The lack of user involvement in the development of public apps may be under-represented however, with development undertaken in-house, potentially without public reporting of the process for proprietary reasons. The public app model involves early deployment and a fast update cycle, which would encourage user involvement.

It's important to note that differences in the composition of the development teams potentially helps to explain variations in aesthetics, functionality, credibility, content and especially, theoretical underpinnings, although evidence for this suggestion has not been identified in the literature to date. Therefore, it's possible that the strategy taken by the funder, be it public or private, may be particularly influential in the final app design.

Effectiveness, impact, evidence-based content and theoretical underpinnings

Apps were perceived as being moderately effective using the MARS score, which was based on perceived impact and existing published evidence. Public apps scored slightly below average, although user reviews suggest high levels of perceived effectiveness, and apps promoting PA alone scored better than apps targeting multiple behaviours.

Individual app components were seen as particularly motivating or as influencing physical activity with reminders, social comparisons and goals. These techniques and features are supported in the behaviour change and physical activity literature as being effective (Epton et al., 2017; Gardner et al., 2015; Howlett et al., 2018; Samdal et al., 2017; Webb et al., 2010).

While most apps included monitoring the emotional consequences of a specific behaviour (BCT 5.4), very few provided information about emotional consequences (BCT 5.6), to suggest that users might experience emotional, as well as physical outcomes, from a behaviour. This reinforces the suggestion made in the introduction to this thesis, that long-term physical health benefits are frequently used in promotional interventions, rather than more immediate, emotive benefits. This is subsequently moderately supported by the presence of BCT 5.1 (Information about health consequences), appearing in more apps (n=7) than BCT 5.6 (n=3).

The prevalence of techniques from the Goals and Planning and Feedback and Monitoring clusters of the taxonomy, would suggest that developers are aware of the evidence in favour of using theory-based self-regulation strategies for behaviour change (Dombrowski et al., 2012; Greaves et al., 2011; Harkin et al., 2016; Michie et al., 2009). Use of these techniques also added to existing evidence for their frequency of use in health and physical activity apps (Edwards et al., 2016; Middelweerd et al., 2014). Despite this, few apps explicitly reported theoretical underpinnings to their apps or for their choice of techniques, mirroring findings in the digital health literature (Bort-Roig et al., 2014; Cowan et al., 2013; Garnett et al., 2018; Zhao et al., 2016).

The links between BCTs and mechanisms of action (MoA), have recently been mapped to determine how BCTs may operate to change behaviours (Carey et al., 2018; Connell et al., 2018; Johnston et al., 2018). Of the 44 BCTs that were used in the included apps, 36 demonstrated links to one of 21 different MoAs (The Theory and Techniques Tool, n.d.). Some of these MoAs have been demonstrated for PA apps previously (Hoj et al., 2017). Those BCTs in clusters related to self-regulation were linked to 12 different MoAs suggesting that the apps have the potential to change behaviours via a range of MoAs whether or not they describe themselves as theory-based.

Despite the existence of national, evidence-based recommendations for health-enhancing physical activity levels (Department of Health and Social Care, 2019; World Health Organisation, 2010), few apps explicitly used or reported them, conforming to existing literature for PA apps (Knight et al., 2015).

Engagement

Apps scored moderately well for engagement, but it was the most poorly scored feature across MARS, excluding perceived impact. Engagement benefited from tailoring and the variety of or optional features, features that have already been reported as acceptable by users. The poorer perception of app engagement overall, may explain why perceived impact was even lower, supporting the assumption that a level of engagement with digital intervention content is required for a MoA to be activated and an effect to occur (Perski et al., 2017; Yardley et al., 2016).

The limited number of user reviews mentioning engagement corroborates the moderate scores from MARS by suggesting users have adhered to the ‘brag and moan’ model of reviews – users only leave a review when they are extremely satisfied or dissatisfied (Hu et al., 2006). Few users were either, it would seem.

Functionality and Security/Privacy

Apps scored well for functionality and those that promoted multiple behaviours, did slightly better than those promoting PA alone. However, user reviews and additional evidence reported mixed evidence of functionality in terms of ease of use and level of effort. In line with the aforementioned 'brag and moan' model (Hu et al., 2006), many users 'moaned' about public app functionality, even with the inclusion of technology experts in development teams, with it being the most frequently cited theme. Complaint types from two papers were used to deductively analyse data from the user reviews. Functionality findings matched three of the types described; errors, crashing/unresponsive apps and network problems (Khalid et al., 2014). However, two other prevalent issues arose in the data: syncing and updates, the latter of which is supported by more recent evidence (McIlroy et al., 2016). While regular updates can be costly for users, it seems they can also anger and drive users away when they impact functionality.

Where public apps fared better than literature-based apps was with respect to security and privacy features. Data protection regulations will have applied to each literature-based research app (e.g. General Data Protection Regulations 2018 (EU), the Privacy Act 1988, 2017 (Australia), Personal Information Protection Action 2011, 2017 (Republic of Korea) and the Health Insurance Portability and Accountability Act 1996 (USA)). Researchers are therefore used to obeying such rules as they, along with ethical practices, form an inherent part of research with participants. It may be why explicit reference to security and privacy features was less extensive for such apps, as it is acknowledged practice within the research domain. Public apps in contrast reported them more consistently. This is reassuring, as a recent review suggested that health apps posed significant privacy and security risks to users (Scott et al., 2016), and users would use an app or share data only if optional, and security and privacy was assured (Dennison et al., 2013; Peng et al., 2016).

Usage and compliance

Positive usage patterns were typically reported when various comparisons were made to other versions of the app, other apps, or over time and some features were used more often than others. Interestingly, these findings appear to refute evidence that suggests app adherence is poor (Jee, 2017) and public app retention could be dropping, although retention has consistently been between 30-40% since 2012 (Iqbal, 2019). The same industry report suggests that retention rates for Android apps drops significantly after a week and continues to fall, but more slowly, in the 90 days since installation. However, as this is for top apps in general and not health apps, it may be that such data is not representative of this domain. Another recent industry review of health app usage suggested that more than 75% of active users opened their apps at least once a week (Kesiraju and Vogels, 2017), although these users may not be representative of the users of the apps included in this thesis, as research studies typically report fixed durations for app use/testing (typically 3-6 months). Given their classification as '*active*', the users described in the industry report were potentially already in the "*acting*" stage of app adoption, described as *"I am currently using an app for that [behaviour] and intend to continue to use it"* (König et al., 2018).

Overall quality

MARS scores for overall quality were moderate to high. However, across all apps, quality as defined by the full 13 criteria was mixed. Although as a group the criteria were all represented, some apps neglected to report some criteria, for example literature-based apps rarely provided sufficient information for functionality evaluation. Others performed poorly on some criteria for example, the Unnamed app scored below average for credibility, aesthetics and engagement, while others excelled - Fitbit achieved a near perfect average score for both aesthetics and engagement. This corroborates previous reviews of digital physical activity interventions (Bort-Roig et al., 2014) and health focused apps showing variation in their quality (Reynoldson et al., 2014). Further analysis would be useful to determine how and if quality criteria inter-relate, beyond the suggestions made in this discussion. Previous research has suggested that use of certain BCTs may be correlated with MARS scores for example (Bardus et al., 2016).

Public apps scored slightly higher than literature-based apps overall, but there were negligible differences between those promoting PA or multiple behaviours and free or paid apps. The last finding adds support to the existing literature on the similarity of the quality of paid and free health apps (Bardus et al., 2016; Bondaronek et al., 2018). This contrasts with evidence that suggests more expensive apps are preferable to cheaper apps (West et al., 2012). Based on data reported in a review of physical activity apps for children, mean scores (calculated specifically for this thesis) also supported the negligible differences in quality between apps targeting PA or multiple behaviours (Schoeppe et al., 2017). Thirteen apps were reported as targeting physical activity and/or sedentary behaviour and scored a mean of 3.48 on overall quality measured by MARS, compared to six apps coded as targeting physical activity and/or sedentary behaviour as well as other behaviours, which scored a mean of 3.97.

6.3 Strengths and limitations

6.3.1 General methods

There are limitations to the methods employed in the two studies. The literature review did not capture conference abstracts, only conference papers. Given the novelty and popularity of apps in health research it’s possible that some relevant apps were missed. In addition, reference lists for systematic reviews and eligible app papers were examined for further relevant manuscripts. There is a suggestion that citation bias can be increased with this method, leading to representation of the same apps repeatedly, but neglecting harder-to-find, unrepresented apps (Higgins and Green, 2011). However, given the limited response from authors who had already published on such apps, the time-frame and scoping carried out before searches were conducted (see 6.3.3) it seems unlikely that sufficient or voluminous additional evidence would have been available.

The public apps and their supporting evidence may have also been subject to some bias. The identification of the final eligible apps took place, by necessity, some time after the initial lists of popular apps were extracted from App Annie. Therefore, given the fast pace of app development and demonstrated frequency of updates, app availability changed during screening, and versions and content may have changed between initial identification and confirmation of eligibility. This also means that while user reviews were extracted from the date the apps were downloaded onwards, in order to capture the most relevant reviews to the versions evaluated, reviews may have been about subsequent versions. This was evidenced for Fitbit where users reported problems with a recent update, but this was not observed during testing. The credibility of the user reviews was acknowledged as a potential source of bias in section 2.6, with functionality appearing to be most at risk from this (see 5.4.15). Although this data may not reflect the functional quality of the apps at the time they were tested and therefore should be interpreted with caution, it does reflect the types of functional challenges users typically experience. The speed of technological development will be a consistent challenge for research in this area. Despite this, attempts were made to generate a *'comprehensive and representative'* sample of apps with the resources available, with both Android and iOS and free and paid apps eligible for evaluation, an acknowledged strength in the app literature (Cowan et al., 2013).

Identifying the additional evidence for eligible apps from searches and websites demonstrated a comprehensive approach to finding data and presenting a reliable, valid and informed interpretation of each app. However, apps were restricted in terms of the potential amount of additional evidence that could be identified, therefore, it is possible that additional papers describing the included apps were available. Only the first 50 relevant papers were screened by a single reviewer to determine eligibility and limited search terms were used, neglecting quality, evaluation or effectiveness terminology. However, contact with authors and developers and searches of app websites aimed to minimise this risk. Given the financial priorities of public app companies, it’s likely that positive findings would be reported publicly and readily available on their websites. As no or limited additional evidence was found via these channels, where it does exist, it may be less than positive. Therefore, additional data for public apps may be biased towards more positive outcomes.

Despite these limitations, the reported studies have successfully captured a comprehensive catalogue of physical activity apps that provide feedback on affect, not just those available in the public domain, but also those developed in an academic context. To date, there has been no published literature that has attempted to examine these types of apps, or examine apps across both evidence domains. Therefore, this thesis is novel in its attempt to combine academic and industry outputs.

6.3.2 Quality criteria and assessment

Only the primary coder and one of four second coders completed online training in coding interventions using the BCT taxonomy (Michie et al., 2013). This, along with differences in the number of BCTs coded for apps reported in additional evidence, suggests that this quality criteria may have limited reliability. Having said that, where BCTs were reported in other papers, there was typically considerable overlap, and differences between the coders in the present studies were minimal. What was challenging however, was the application of the BCT taxonomy to digital interventions. This resulted in the need to define how certain app content would be coded, for example badges or point-based rewards. Although the taxonomy is lengthy, and extensively used in the behaviour change literature, it currently does not include examples from digital interventions, which would be of great value in the future. Relying on described content for literature-based apps was also challenging. As with many traditional interventions, as well as previous reviews of apps, despite the existence of reporting guidelines for app studies (BinDhim et al., 2014) and intervention content (Hoffmann et al., 2014), reporting left much to be desired (Dombrowski et al., 2007; Zhao et al., 2016).

Creation of quality criteria for the purposes of this thesis was a pragmatic decision. Although the criteria had face validity given that they were compiled from existing quality assessment literature, they were not independently validated. From the available data, it was clear that apps had not had their quality so comprehensively assessed previously, a strength of the thesis. However, this meant that predictive and concurrent validity of the tools used cannot be established, as criterion have rarely been assessed previously by other standardised measures.

The MARS was particularly valuable in assessing content and added substantial amounts of previously missing or limited data for both sets of apps, making between-app comparisons more feasible. However, MARS is subjective and dependent on access to the app or its description, which was limited for literature-based apps (Stoyanov et al., 2015). In addition, despite recommendations that a variety of study designs should be employed to evaluate app effectiveness (Michie et al., 2017), MARS appears biased towards RCTs. Apps could score higher for Evidence Base (item 19) if they reported one or more RCTs favouring their effectiveness. MARS is also questionable with respect to its initial Engagement item, Entertainment (item 1). This item suggests gamification could be used to make the app more entertaining, but doesn’t define gamification. Gamification can be operationalised in multiple ways, including provision of badges or trophies, leaderboards, points and levels, challenges and quests, or social engagement loops and onboarding (Miller et al., 2014).

Judgements were made with respect to differences within the included apps, specifically between those that targeted PA alone or alongside other behaviours and those that were from the literature base compared to public apps. The small sample size meant that statistical tests of differences between MARS scores were not performed. Therefore these findings should be interpreted with caution. However, given that these apps have not been explored before, and that more apps could not be found, this data should serve as the basis for further fine-grained analysis of potential differences.

One potential limitation to the quality assessment performed in this thesis, was the influence of the researcher. Reflexivity is an acknowledged characteristic of qualitative and therefore mixed-methods research that includes qualitative approaches. It allows researchers to consider the bias they may bring to qualitative research, but it can be both a challenge and an opportunity (Finlay, 2002). One challenge relating to reflexivity was the completion of MARS. Having spent more than ten years in health-based intervention research, it was a struggle to disassociate from a critical perspective on intervention content, mode of delivery and evidence, to become merely an app user. In addition, double coders were also health intervention experts. Therefore, scores may have been biased towards more negative ends of the scale. The same could be said for the thematic analysis of user reviews – previous knowledge of quality criteria is difficult to disconnect from, in order to allow inductive analysis to occur. However, discussions of the analysis with the supervisory team, who were less familiar with the criteria, suggested themes had face validity. In addition, the inclusion of user reviews in the thesis allowed a degree of reflexivity. It offered a way to include user voices, with users then acting as co-researchers in the assessment of the quality of the apps. One weakness of this thesis could be that app developers' voices were not included, as they are also a key part of the quality assessment process.

6.3.3 Data set

The lack of available apps for those from the literature was an expected limitation, which resulted in both known missing data and potentially unknown missing data. Although descriptions of literature-based apps were extensive in some cases, the information provided was not sufficient to perform a full, independent evaluation of the content of the app. Speculation made in section 3.4.7.4.1 that the poor availability of such apps may have been due to funding issues, among others, was supported by author correspondence reporting just that (appendix 23). Correspondence also suggested that academics struggled to get developers to commit to maintaining the app beyond its initial release (appendix 24). For another app, a modular evaluation approach had been taken and, therefore, the parts of the app were being used elsewhere, but not in the same configuration as the described app (appendix 25). Therefore, data, especially for MARS and BCTs, was based on information that was limited and sometimes held outside of the apps, such as affiliated websites. As such, findings, recommendations and conclusions are based on limited information and information that is not necessarily available in the apps themselves.

The ratio of screened to available public apps was also an expected limitation. The volume of health apps available in stores, availability of compatible devices, and costs, precluded exhaustive identification and evaluation. In addition, the popularity and availability of apps fluctuated daily, as demonstrated by daily changes in top app lists on App Annie. Therefore, the current findings are limited in their longevity. This is the case for all technology however, and only a paradigm shift in the way research is conducted can address this issue fully.

Only English language apps were included. There is evidence that additional, foreign language apps may have been relevant for inclusion, such as Leefplezier (Blaauw et al., 2014), however only a Dutch version was available. Future work in this area will require support from multi-lingual collaborators.

Revisions to the eligibility criteria saw the inclusion of apps that promoted multiple behaviours in addition to physical activity. This criterion was added after searches had been run, suggesting that eligible apps may be missing from the current data set. However, extensive scoping of the literature took place over a period of at least two years prior to carrying out the final searches. During that time, no additional apps that fit the eligibility criteria were identified. One was identified that targeted an incorrect population (Morandi and Serafin, 2007) while some studies provided feedback on affect, but didn’t use an app (e.g. Tanaka et al., 2016) and many were found that measured affect, but failed to provide feedback on it (e.g. Hassandra et al., 2017). The inclusion of such apps made coding more challenging, as BCTs sometimes targeted other behaviours, but were included alongside physical activity promoting strategies, calling into question whether or not these BCTs should be reported. These apps also complicated any ability to infer the impact of feedback on affect, or any BCT, on users' physical activity levels. However, as effectiveness and mechanism of effect were not explicitly investigated in this thesis, this was of less importance than attempting to be comprehensive in capturing apps that provided feedback on affect.

6.3.4 Generalisability and transferability of findings

Supporting evidence suggests that many of the findings relating to quality are likely to be generalisable to other physical activity apps and active healthy adult populations. However, ecological, alternative population and temporal transferability are less definitive (Teddlie and Tashakkori, 2009). The findings may not transfer to apps created in under-developed countries, or to individuals at risk of inactivity and who are not already active. In addition, as previously mentioned, the pace of app development suggests that at least some findings may not be transferable over time. For example, improvements in accuracy and reliability of devices are likely to continue, meaning that issues of credibility related to this facet of quality may disappear. International developments in security and safety requirements such as the recent update to the Data Protection Act (DPA 2018), as well as evidence standards, such as the recent framework for digital health technologies (NICE, 2019), will mean improvements should be forthcoming in these quality areas. In fact, some facets of quality reported in the current thesis may become obsolete as technology progresses. Although, numerous reviews of public and literature-based apps highlight the fact that there have been continuing concerns with respect to evidence-based content, theoretical underpinnings and usage, irrespective of the target behaviour or population (Bardus et al., 2016; Breton et al., 2011; Cowan et al., 2013; Jee, 2017; Pagoto et al., 2013).

6.4 Implications and recommendations

In the introduction, a number of problems were posed as potential issues for physical activity interventions. These were poor reach to those most in need, ineffective content and poor adoption, fidelity and a subsequent lack of focus on maintenance of adopted behaviours. Apps were proposed to have the potential to fill these gaps.

Findings from the present studies suggest that while popular and acceptable, it’s unclear whether those most in need of physical activity for health are being reached by these apps. Unfortunately, as the literature demonstrates, such apps may be more favourable to younger, more educated, already active, conscientious individuals interested in self-quantification (Bielik et al., 2012; Bol et al., 2018; Maltseva and Lutz, 2018). However, the fact that the majority of the apps were free is encouraging and reduces issues of accessibility. In addition, the apps frequently captured walking (or step counts), a popular and accessible activity for many at-risk groups (Salmon et al., 2003; Siegel et al., 1995). Physical activity apps that include feedback on affect have characteristics that increase accessibility for at-risk populations, but more explicit targeting would be advised.

Findings from the present studies suggest that evidence-based content in these apps is prevalent in terms of BCTs and those specifically associated with increasing PA. However, theoretical underpinnings and inclusion of evidence-based guidelines for PA were severely lacking. The latter is of particular concern as those more at risk from inactivity have demonstrated that not only are they unconvinced of the benefits of PA, but are also less likely to be aware of there being any guidelines at all (Knox, Esliger, et al., 2013; Murray, 2006). Physical activity apps that include feedback on affect include evidence-based content, but could include additional content that targets specific barriers for at-risk groups, such as information about PA recommendations and benefits, cost-efficient PA, and suggestions for safe and culturally appropriate activities (Seefeldt et al., 2002).

Findings from the present studies suggest that while included public apps are popular (they all had multiple downloads and high user ratings), engagement is questionable due to limited data, long-term usage is unclear and effort is still required to ensure fidelity before effectiveness can be tested.

In addition to these gaps, findings from the present studies (the limited use of BCT 5.6 Information on emotional consequences and the rare focus on PA-contingent affect) also imply that feedback on affect, as a way to facilitate physical activity behaviour change, is lacking in apps. If positive affect is not anticipated due to poor or limited attempts to draw users' attention to the affective benefits of PA, then users may be less likely to change their behaviours (Conner, 2018; Dunton and Vaughan, 2008). Indeed, anticipated affect (AA) could map onto a MoA, such as '*beliefs about consequences'*, which has been linked to the BCT 5.6 Providing information on emotional consequences, and BCT 5.2 Salience of consequences (Carey et al., 2018; Connell et al., 2018; Johnston et al., 2018). These BCTs were rarely used in the present included apps. Similarly, due to these missing components it's unlikely that the lack of belief in the affective benefits of PA was addressed by the current apps. This is important, as this gap has been identified in those more at risk of being inactive (Murray, 2006), although there is also evidence to suggest that for those that are already active negative affect can be motivating (Speranzini, 2015). Unfortunately, the latter are not the population that need most attention.

While the characteristics of the included apps have now been established (with the exception of security features, developers and target populations), evidence for their quality has been decidedly mixed. In combination with concerns raised above, additional app development is required before questions of effectiveness of these types of apps can be addressed. Therefore, in their present form, PA apps that provide feedback on affect are not currently the solution to poor PA levels.

Given the findings, implications, and acknowledged limitations of the studies and data, a number of recommendations can be made for future investigation. These recommendations are targeted at researchers, app developers and funders, rather than health care practitioners or policy makers, due to the emergent state of the literature.

6.4.1 Recommendations for research and researchers

1) Physical activity apps that include feedback on affect need to be developed with each aspect of quality taken into consideration.
2) Such apps should be developed in collaboration with behaviour change, health and technology experts and users.
3) Such apps should make efforts to accommodate and target those at most risk of inactivity, both in terms of content and importantly, marketing, to ensure meaningful reach and uptake.
4) Such apps should be investigated for their mechanisms of effect/action, efficacy and effectiveness, especially in relation to feedback on affect.
5) Such apps should include evidence-based content that targets relevant mechanisms of change.
6) Researchers should make more information available about apps they have developed to allow future evaluations. Preferably the app would be archived in some way. Android Application Package (APK) files could be stored using the Open Science Framework (Centre for Open Science, 2011). Reporting standards for interventions would also be used in addition to specific digital-based data such as release dates, security and technical features and developers.
7) Applications for funding should include consideration of how apps will be maintained or preserved for future evaluation.
8) Researchers should seek to develop a feasible and comprehensive standardised quality assessment tool for apps.
9) While awaiting a gold standard for quality assessment, researchers should consider using the Mobile App Rating Scale (Stoyanov et al., 2015) in the first instance to evaluate a breadth of quality criteria, but also consider additional criteria reported in this thesis.
10) Researchers should use established usability and user-experience tests as well as traditional research methods to assess quality.
11) Many correlational ecological momentary assessment studies using smartphones were excluded from the review, due to their investigation of the nature of the relationship between activity and affect. Researchers should now investigate the relationship using

experimental methods with a focus on changing physical activity levels using an app.

6.4.2 Recommendations for app developers

1) Commercial developers should consider providing transparent development procedures.
2) Developers should consider reporting the evidence-base for their products.
3) Developers should consider providing more informative app descriptions.

6.4.3 Recommendations for funders

1) Funders should consider resource requirements for high quality app development and maintenance.
2) Funders should encourage collaboration between researchers and app developers as well as the involvement of users.

6.4.4 Future research

The next phase of this research could be to develop a paper prototype of a physical activity app that provided feedback on affect that attempts to fill the identified granular quality gaps. Testing and refining the vision for an app are recommended prior to spending precious funding on programming (Roth et al., 2014). Paper prototyping is a cost-effective, resource efficient, user-centred design method, familiar and acceptable to developers. It is used in digital technology development and allows early usability testing and rapid refinements to take place (Grady, 2000; Peters et al., 2018). Development would involve collaboration between users, developers and academics. The prototype would aim to target inactive individuals. Once quality was established, a working prototype would be programmed and further quality and feasibility assessments performed, followed by piloting and modelling of process and outcomes, as recommended by the Medical Research Council

(Craig et al., 2008). This would culminate in a test of effectiveness using a modular approach, as recommended in the literature, in the first instance.

High quality, effective apps are just the beginning for this domain of physical activity behaviour change. With the growing capabilities of smartphones and popularity of just-in-time-interventions, there is scope for interventions to perform signal-detection for low affect, for example by identifying slow speech cadence, or by detecting low activity levels in activity-compatible contexts, and subsequently promote physical activity (Hardeman et al., 2019; Kanjo et al., 2018; Pentland et al., 2013). The pace of technological change may be an ongoing challenge, but there is still valuable learning to be acquired for researchers with respect to the quality and mechanisms of effect for physical activity apps.

6.5 Conclusions

Physical activity apps that include feedback on affect already exist but there are few of them that capture and provide feedback specifically on PA-contingent affect. The characteristics of the apps reveal that they rarely target or report to target, those most at risk of inactivity. However, their cost and the inclusion of walking as an activity does suggest that these apps are accessible and acceptable to such groups. Quality of these apps is mixed, but promising in some areas. However, the information available to allow a quality assessment to be performed was limited. Future research should include developing higher quality apps, using feedback on affect explicitly linked to physical activity and targeting those at risk, before effectiveness is rigorously tested. Researchers, app developers and funders need to consider the longevity of their apps. The former two should be encouraged to collaborate (alongside users) in order to develop both aesthetically pleasing, engaging and functional, as well as theory and evidence based apps. Funders should consider the cost of high quality development and importantly, maintenance of apps, to ensure resources aren’t wasted and quality and effectiveness can be assured.

7. References

Aaltonen S, Kujala UM and Kaprio J (2014) Factors behind leisure-time physical activity behavior based on Finnish Twin Studies: The role of genetic and environmental influences and the role of motives. *BioMed Research International*: 1–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3997869&tool=pmcentrez&rendertype=abstract.

Aaltonen S, Rottensteiner M, Kaprio J, et al. (2014) Motives for physical activity among active and inactive persons in their mid-thirties. *Scandinavian Journal of Medicine & Science in Sports* 24(4): 727–735.

Achterkamp R, Hermens HJ and Vollenbroek-Hutten MMR (2015) The influence of success experience on self-efficacy when providing feedback through technology. *Computers in Human Behavior*, Elsevier Ltd 52: 419–423. Available from: http://dx.doi.org/10.1016/j.chb.2015.06.029.

Agicent (2017) How much does it cost to maintain an app? *Mobile App Design & Development, Mobile app basics*. Available from: https://www.agicent.com/blog/how-much-does-it-cost-to-maintain-an-app/.

Alexander GL, McClure JB, Calvi JH, et al. (2010) A randomized clinical trial evaluating online interventions to improve fruit and vegetable consumption. *American Journal of Public Health* 100(2): 319–326.

Ali R, Siddiqi MH, Lee S, et al. (2015) KARE: A hybrid reasoning approach for promoting active lifestyle. *ACM IMCOM 2015 - Proceedings*: 0–4.

Altenhoff B, Vaigneur H and Caine K (2015) One step forward, two steps back: The key to wearables in the field is the app. In: *Pervasive Health, May 20-23, Istanbul, Turkey*.

Anderson K, Burford O and Emmerton L (2016) App chronic disease checklist: Protocol to evaluate mobile apps for chronic disease self-management. *JMIR Research Protocols* 5(4): e204.

Annesi JJ and Westcott WL (2007) Relations of physical self-concept and muscular strength with resistance exercise-induced feeling state scores in older women. *Perceptual & Motor Skills* 104(1): 183–190. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=24774213&site=ehost-live.

Anokye NK, Trueman P, Green C, et al. (2012) Physical activity and health related quality of life. *BMC Public Health*, BMC Public Health 12(1): 1. Available from: BMC Public Health.

Antón AM and Rodríguez BR (2016) Runtastic PRO app: an excellent all-rounder for logging fitness. *British Journal of Sports Medicine* 50(11): 705–706.

App Annie (2010). Available from: https://www.appannie.com/en/ (accessed 15 January 2019).

App Annie Support (2018) Personal communication with App Annie Support. 23 July.

Apple Inc (2019) App Store Connect Help. Available from: https://help.apple.com/app-store-connect/#/devcdda55918 (accessed 29 July 2019).

AppTornado GmbH (2010) AppBrain. Available from: https://www.appbrain.com/ (accessed 15 January 2019).

Arambepola C, Ricci-Cabello I, Manikavasagam P, et al. (2016) The impact of automated brief messages promoting lifestyle changes delivered via mobile devices to people with type 2 diabetes: A systematic literature review and meta-analysis of controlled trials. *Journal of Medical Internet Research* 18(4): 1–17.

Armitage A (2007) Mutual research designs: Redefining mixed methods research design. In: *British Educational Research Association Annual Conference*, London.

Bandura A (1986) *Social foundations of thought and action*. Englewood Cliffs. New Jersey: Prentice-Hall International.

Bandura A (1991) Social Cognitive Theory of Self-Regulation. *Organizational behavior and human decision processes* 50(2): 248–287.

Bankhead P (2017) Welcome to Google Developer Day. In: *Game Developer Conference*, Android Developers Blog. Available from: https://android-developers.googleblog.com/2017/02/welcome-to-google-developer-day-at-game.html.

Bardus M, van Beurden SB, Smith JR, et al. (2016) A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management. *International Journal of Behavioral Nutrition and Physical Activity* 13(1): 35. Available from: http://www.ijbnpa.org/content/13/1/35.

Baretta D, Perski O and Steca P (2018) Exploring users' experiences of the uptake and adoption of physical activity apps: longitudinal qualitative study. *JMIR mHealth and uHealth* 7(2): e11636.

Bauer C and Kriglstein S (2016) Analysis of motivation strategies in running tracking applications. In: *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*, Brussels, Belgium, December 11-13, 2015, pp. 73–79.

Bauman AE, Reis RS, Sallis JF, et al. (2012) Correlates of physical activity: Why are some people physically active and others not? *The Lancet*, Elsevier Ltd 380(9838): 258–271. Available from: http://dx.doi.org/10.1016/S0140-6736(12)60735-1.

Baumeister RF, Vohs KD, DeWall CN, et al. (2007) How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review* 11(2): 167–203.

Beltrán-Carrillo VJ, Jiménez-Loaisa A, Alarcón-López M, et al. (2019) Validity of the "Samsung Health" application to measure steps: A study with two different samsung smartphones. *Journal of Sports Sciences*, Routledge

37(7): 788–794. Available from: https://doi.org/10.1080/02640414.2018.1527199.

Bentley F, Tollmar K and Stephenson P (2013) Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *Tochi* 20(5): 1–27. Available from: http://dl.acm.org/citation.cfm?id=2503823.

Bickmore TW, Mauer D and Brown T (2009) Context awareness in a handheld exercise agent. *Persuasive mobile computing* 5(3): 226–235.

Biddle S (1992) Adherence to physical activity and exercise. In: Norgan N (ed.), *Physical Activity and Health*, Cambridge: Cambridge University Press, pp. 170–189.

Biddle S (1995) *European Perspectives on Exercise and Sport Psychology*. Biddle SJH (ed.), Leeds: Human Kinetics.

Biddle SJH and Mutrie N (2008) Physical activity: a feel good effect? 2nd Editio. In: *Psychology of Physical Activity: determinants, well-being & interventions*, Abingdon: Routledge.

Bielik P, Tomlein M, Krátky P, et al. (2012) Move2Play: An innovative approach to encouraging people to be more physically active. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12* (January 2014): 61. Available from: http://dl.acm.org/citation.cfm?doid=2110363.2110374.

Biesta G (2010) Pragmatism and the philosophical foundations of mixed methods research. 2nd Editio. In: Tashakkori A and Teddlie C (eds), *SAGE Handbook of Mixed Methods in Social and Behavioural Research*, London: Sage Publications Ltd, pp. 95–118.

Billiet W and Vanden Bûssche B (2016) Quality criteria for health apps: a systematic review. KU Leuven.

BinDhim NF and Trevena L (2015) There's an app for that: A guide for Healthcare Practitioners and researchers on smartphone technology.

Online Journal of Public Health Informatics 7(2): 1–12. Available from: http://journals.uic.edu/ojs/index.php/ojphi/article/view/5522.

BinDhim NF, Hawkey A and Trevena L (2014) A systematic review of quality assessment methods for smartphone health apps. *Telemedicine journal and e-health: the official journal of the American Telemedicine Association* 21(2): 1–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25469795.

Blaauw F, Jonge P De and Aiello M (2014) Leefplezier: Personalized well-being. *IEEE Intelligent Informatic Bulletin* 15(1).

Bol N, Helberger N and Weert JCM (2018) Differences in mobile health app use: A source of new digital inequalities? *The Information Society*, Taylor & Francis 34(3): 183–193. Available from: https://www.tandfonline.com/doi/full/10.1080/01972243.2018.1438550.

Bondaronek P (2017) Personal communication with Paulina Bondaronek. 7 March.

Bondaronek P, Alkhaldi G, Slee A, et al. (2018) Quality of publicly available physical activity apps: review and content analysis. *JMIR mHealth and uHealth* 6(3): e53. Available from: http://mhealth.jmir.org/2018/3/e53/.

Bonham T, Pepper G V. and Nettle D (2018) The relationships between exercise and affective states: a naturalistic, longitudinal study of recreational runners. *PeerJ* 6: e4257. Available from: https://peerj.com/articles/4257.

Bort-Roig J, Gilson ND, Puig-Ribera A, et al. (2014) Measuring and influencing physical activity with smartphone technology: A systematic review. *Sports Medicine* 44(5): 671–686.

Boudreaux ED, Waring ME, Hayes RB, et al. (2014) Evaluating and selecting mobile health apps: strategies for healthcare providers and healthcare organizations. *Translational Behavioral Medicine* 4(4): 363–371.

Bouts AM, Brackman L, Martin E, et al. (n.d.) The accuracy and validity of iOS-based heart rate apps during moderate to high intensity exercise. *International journal of exercise science* 11(7): 533–540. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29541341%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5841680.

Brainard J, Wilsher SH, Salter C, et al. (2016) Methodological review: Quality of randomized controlled trials in health literacy. *BMC Health Services Research*, BMC Health Services Research 16(1): 1–10. Available from: http://dx.doi.org/10.1186/s12913-016-1479-2.

Brand R and Antoniewicz F (2011) Affective evaluations of exercising: The role of automatic-reflective evaluation discrepancy. *International Journal of Sport Nutrition and Exercise Metabolism* 32(December): 1–44.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3: 77–101.

Brennan RL and Prediger DJ (1981) Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41(3): 687–699.

Breton ER, Fuemmeler BF and Abroms LC (2011) Weight loss-there is an app for that! But does it adhere to evidence-informed practices? *Translational Behavioral Medicine* 1(4): 523–529.

Brindal E, Hendrie GA and Freyne J (2016) Combining persuasive technology with behavioral theory to support weight maintenance through a mobile phone app: Protocol for the MotiMate app. *JMIR Research Protocols* 5(1): e5–e5. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=26747725&site=ehost-live.

Brindal E, Hendrie G and Noakes M (2018) A smartphone application designed to support weight loss maintenance and wellbeing: Results of a randomized trial for the MotiMate app (PREPRINT).

Brinkel J, Krämer A, Krumkamp R, et al. (2014) Mobile phone-based mHealth approaches for public health surveillance in sub-Saharan

Africa: a systematic review. *International journal of environmental research and public health* 11(11): 11559–11582.

British Sociological Society (2017) Statement of Ethical Practice Annexe Ethics Guidelines and Collated Resources for Digital Research BSA Statement of Ethical Practice Annexe 2 Ethics Guidelines and Collated Resources for Digital Research. Available from: www.britsoc.co.uk.

British Standards Institution (2015) Health and wellness apps - Quality criteria across the life cycle - Code of practice (PAS 277:2015).: 36. Available from: http://shop.bsigroup.com/upload/271432/PAS 277 (2015)bookmarked.pdf.

Brooke J (1996) SUS: a 'quick and dirty'usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, et al. (eds), *Usability Evaluation in Industry*, London: Taylor & Francis. Available from: https://books.google.co.uk/books?hl=en&lr=&id=IfUsRmzAqvEC&oi=fnd&pg=PA189&dq=System+Usability+Scale+(SUS%3B+Brooke,+1996&ots=GamyG7ji5l&sig=xRYlM5fmLi-wnS5ixskcLabBHpU#v=onepage&q=System Usability Scale (SUS%3B Brooke%2C 1996&f=false.

Bull FC, Kreuter MW and Scharff DP (1999) Effects of tailored, personalized and general health messages on physical activity. *Patient education and counseling* 36(2): 181–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10223022.

Buman MP, Epstein DR, Gutierrez M, et al. (2016) BeWell24: development and process evaluation of a smartphone “app” to improve sleep, sedentary, and active behaviors in US Veterans with increased metabolic risk. *Translational Behavioral Medicine* 6(3): 438–448.

Burns P, Lueg C and Berkovsky S (2012) Activmon: encouraging physical activity through ambient social awareness. In: *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, Austin, Texas, USA, May 5-10, 2012, pp. 2363–2368. Available from: http://dl.acm.org/citation.cfm?id=2223803.

Business of Apps (2016) *App Statistics Report*. Available from: http://www.businessofapps.com/data/app-statistics/.

Byrne A and Byrne DG (1993) The effect of exercise on depression, anxiety and other mood states: A review. *Journal of Psychosomatic Research* 37(6): 565–574.

Byrt T, Bishop J and Carlin JB (1993) Bias, prevalence and kappa. *J. Clinical Epidemiology* 46(5): 423–429.

Caldeira C, Chen Yu, Chan L, et al. (2017) Mobile apps for mood tracking: an analysis of features and user reviews. *AMIA 2017 Annual Symposium*. Available from: http://www.ics.uci.edu/~claram/pdf/MoodTracking-AMIA17.pdf.

Carels RA, Coit C, Young K, et al. (2007) Exercise makes you feel good, but does feeling good make you exercise? an examination of obese dieters. *Journal of sport & exercise psychology* 29(6): 706–722.

Carey RN, Connell LE, Johnston M, et al. (2018) Behavior change techniques and their mechanisms of action: A synthesis of links described in published intervention literature. *Annals of Behavioral Medicine*: 1–15.

Carroll JK, Moorhead A, Bond R, et al. (2017) Who uses mobile phone health apps and does use matter? A secondary data analytics approach. *J Med Internet Res* 19(4): e125.

Carson D, Gilmore A, Perry C, et al. (2001) *Qualitative Marketing Research*. SAGE (ed.), London. Available from: https://uk.sagepub.com/en-gb/eur/qualitative-marketing-research/book208811#preview.

Case MA, Burwick HA, Volpp KG, et al. (2015) Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Journal of American Medical Association* 313(6): 625. Available from: http://jama.jamanetwork.com/article.aspx?articleid=2108876.

Caspersen CJ, Powell KE and Christenson GM (1985) Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports (Washington, D.C. : 1974)* 100(2): 126–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3920711%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1424733.

Cavill N and Bauman A (2004) Changing the way people think about health-enhancing physical activity: Do mass media campaigns have a role? *Journal of Sports Sciences* 22(8): 771–790.

Centre for Open Science (2011) Open Science Framework. Available from: https://osf.io/ (accessed 2 August 2019).

Centre for Reviews and Dissemination (2009) *Systematic reviews: CRD's guidance for undertaking reviews in health care*. University of York: Centre for Reviews and Dissemination.

Chau M and Reith R (2019) *Smartphone market share*. *IDC Device Market Trends*. Available from: https://www.idc.com/promo/smartphone-market-share/os.

Chekroud SR, Gueorguieva R, Zheutlin AB, et al. (2018) Association between physical exercise and mental health in 1·2 million individuals in the USA between 2011 and 2015: a cross-sectional study. *The Lancet Psychiatry*, Elsevier Ltd 5(9): 739–746. Available from: http://dx.doi.org/10.1016/S2215-0366(18)30227-X.

Cho H and Salmon CT (2007) Unintended effects of health communication campaigns. *Journal of Communication* 57(2): 293–317.

Cho Y-M, Lee S, Islam SMS, et al. (2018) Theories applied to m-Health interventions for behavior change in low- and middle-income countries: A systematic review. *Telemedicine and e-Health* 24(11): tmj.2017.0249. Available from: http://online.liebertpub.com/doi/10.1089/tmj.2017.0249.

Choe EK, Lee NB, Lee B, et al. (2014) Understanding quantified-selfers' practices in collecting and exploring personal data. *Conference on Human Factors in Computing Systems - Proceedings*: 1143–1152.

Clark AM (1998) The qualitative-quantitative debate: Moving from positivism and confrontation to post-positivism and reconciliation. *Journal of Advanced Nursing* 27(6): 1242–1249.

Cleland CL, Tully MA, Kee F, et al. (2012) The effectiveness of physical activity interventions in socio-economically disadvantaged communities: a systematic review. *Preventive medicine*, Elsevier Inc. 54(6): 371–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22521997 (accessed 4 June 2013).

Cleland I, Han M, Nugent C, et al. (2013) Mobile based prompted labeling of large scale activity data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8277 LNCS: 9–17.

Clow A and Edmunds S (2014) Biological foundations of physical activity on mental health. In: Clow A and Edmunds S (eds), *Physical activity and mental health*, Leeds: Human Kinetics, pp. 8–14.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and psychological measurement* (1): 37–46.

Cohen J (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4): 213–220.

Colineau N and Paris C (2011) Motivating reflection about health within the family: The use of goal setting and tailored feedback. *User Modeling and User-Adapted Interaction* 21(4–5): 341–376.

Connell LE, Carey RN, de Bruin M, et al. (2018) Links between behavior change techniques and mechanisms of action: An expert consensus study. *Annals of Behavioral Medicine*: 1–13.

Conner Mark (2018) Experiential attitude and anticipated affect. In: Williams D, Rhodes RE, and Conner M (eds), *Affective Determinants of Health-Related Behavior*, Oxford University Press.

Consolvo S, Mcdonald DW, Toscos T, et al. (2008) Activity sensing in the wild: A field trial of UbiFit Garden. In: *CHI 2008*, pp. 1797–1806.

Cooney G, Dawn K, Greig C, et al. (2013) Exercise for depression (Review). *The Cochrane Library* (9).

Cooper AR, Goodman A, Page AS, et al. (2015) Objectively measured physical activity and sedentary time in youth: The International children's accelerometry database (ICAD). *International Journal of Behavioral Nutrition and Physical Activity* 12(1): 1–10. Available from: http://dx.doi.org/10.1186/s12966-015-0274-5.

Costantino TE (2012) Constructivism. In: Given LM (ed.), *The Sage encyclopedia of qualitative research methods*, Thousand Oaks: SAGE Publications, pp. 318–319.

Coughlin SS, Whitehead M, Sheats JQ, et al. (2016) A review of smartphone applications for promoting physical activity. *Jacobs J Community Med* 2(1).

Cowan LT, Van Wagenen SA, Brown BA, et al. (2013) Apps of steel: are exercise apps providing consumers with realistic expectations?: a content analysis of exercise apps for presence of behavior change theory. *Health Education & Behavior* 40(2): 133–9. Available from: http://heb.sagepub.com/content/early/2012/09/17/1090198112452126.abstract (accessed 29 October 2014).

Craig P, Dieppe P, Macintyre S, et al. (2008) Developing and evaluating complex interventions: The new Medical Research Council guidance. *Bmj* 337(7676): 979–983.

Crawford BM (2013) Wellness for free: website and app review. *Journal of the American Chiropractic Association (Arlington, Va.)* Jan-Feb: 18–22.

Cresswell JW (2003) *Research design: Qualitative, quantitative, and mixed methods approaches*. 2nd ed. London: Sage Publications Ltd.

Cresswell JW (2007) *Qualitative inquiry and research design: Choosing* *among five approaches*. 2nd Editio. London: SAGE Publications.

Cresswell JW, Plano-Clark VL, Gutmann ML, et al. (2003) Advanced mixed methods research designs. *Handbook of Mixed Methods in Social and Behavioral Research*: 209–240. Available from: http://www.sagepub.com/upm-data/19291_Chapter_7.pdf.

Crotty M (1998a) Introduction: the research process. In: *The foundations of social research: Meaning and perspective in the research process*, London: SAGE Publications Ltd, p. 256.

Crotty M (1998b) *The foundations of social research: Meaning and perspective in the research process*. London: Sage Publications Ltd.

Cunningham M (2009) More than just the Kappa coefficient: A program to fully characterize inter-rater reliability between two raters. *SAS Global Forum (Sgf) 2009*: 1–7. Available from: http://support.sas.com/resources/papers/proceedings09/242-2009.pdf.

Darby BG, Zhang LA, Owen Y, et al. (2016) Mood alteration after 15 minutes of preferred intensity exercise: examining heart rate, perceived exertion, and enjoyment. *Journal of Sport Behaviour*: 1–13.

De Bruijn GJ, Verkooijen K, de Vries NK, et al. (2012) Antecedents of self identity and consequences for action control: An application of the theory of planned behaviour in the exercise domain. *Psychology of Sport and Exercise*, Elsevier Ltd 13(6): 771–778. Available from: http://dx.doi.org/10.1016/j.psychsport.2012.05.008.

Deline MB, Baumer EPS and Gay G (2012) Normative communication processes and associated emotion in mobile health groups. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion - CSCW ’12*: 75. Available from: http://dl.acm.org/citation.cfm?doid=2141512.2141546.

Deloitte (2016) There’s no place like phone - Global Mobile Consumer Survey 2016. Available from: https://www.deloitte.co.uk/mobileuk/assets/pdf/Deloitte-Mobile-

Consumer-2016-There-is-no-place-like-phone.pdf.

Dennison L, Morrison L, Conway G, et al. (2013) Opportunities and challenges for smartphone applications in supporting health behavior change: Qualitative study. *Journal of Medical Internet Research* 15(4): 1–12.

Denscombe M (2010) *The Good Research Guide for small-scale social research project*. Available from: https://www.researchgate.net/file.PostFileLoader.html?id=582a0dbf217e20276533f5a5&assetKey=AS%3A428404664213506%401479151039119.

Department of Health and Social Care (2019) *UK Chief Medical Officers ' Physical Activity Guidelines*. Available from: https://www.gov.uk/government/publications/physical-activity-guidelines-uk-chief-medical-officers-report.

Desharnais R, Jobin J, Côté C, et al. (1993) Aerobic exercise and the placebo effect: a controlled study. *Psychosomatic medicine* 55(15): 149–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8475229.

Ding D, Lawson KD, Kolbe-Alexander TL, et al. (2016) The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet*, Elsevier Ltd 388(10051): 1311–1324. Available from: http://dx.doi.org/10.1016/S0140-6736(16)30383-X.

Direito A, Dale LP, Shields E, et al. (2014) Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? *BMC Public Health* 14(1): 646. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4080693&tool=pmcentrez&rendertype=abstract (accessed 4 September 2014).

Dishman RK and Buckworth J (1996) Increasing physical activity: a quantitative synthesis. *Medicine & Science in Sports & Exercise* 28(6): 706–719. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage

&an=00005768-199606000-00010 (accessed 18 February 2014).

Dishman RK, Sallis JF and Orenstein DR (1984) The determinants of physical activity and exercise. *Public Health Reports* 100(2): 158–171.

Dogtiew A (2019) App stores list (2018). *Business of Apps*. Available from: https://www.businessofapps.com/guide/app-stores-list/ (accessed 29 July 2019).

Doherty A, Jackson D, Hammerla N, et al. (2017) Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE* 12(2): 1–14.

Dombrowski SU, Sniehotta FF, Avenell A, et al. (2007) Current issues and future directions in psychology and health: Towards a cumulative science of behaviour change: Do current conduct and reporting of behavioural interventions fall short of best practice? *Psychology and Health* 22(8): 869–874.

Dombrowski SU, Sniehotta FF, Avenell A, et al. (2012) Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. *Health Psychology Review* 6(August 2015): 7–32.

Donmoyer R (2012) Quantitative research. In: Given LM (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, Thousand Oaks: SAGE Publications. Available from: http://rcnpublishing.com/doi/abs/10.7748/ns2013.06.27.43.59.s52.

Doshi A, Patrick K, Sallis JF, et al. (2003) Evaluation of physical activity web sites for use of behavior change theories. *Ann Behav.Med* 25(2): 105–111. Available from: pm:12704012.

Duncan MJ, Vandelanotte C, Caperchione C, et al. (2012) Temporal trends in and relationships between screen time, physical activity, overweight and obesity. *BMC Public Health* 12(1): 1.

Dunton GF and Vaughan E (2008) Anticipated affective consequences of physical activity adoption and maintenance. *Health Psychology* 27(6): 703–710.

Dunton GF, Liao Y, Intille SS, et al. (2011) Investigating children’s physical activity and sedentary behavior using ecological momentary assessment with mobile phones. *Obesity*, Nature Publishing Group 19(6): 1205–1212. Available from: http://dx.doi.org/10.1038/oby.2010.302/nature06264.

Dunton GF, Liao Y, Intille S, et al. (2015) Momentary assessment of contextual influences on affective response during physical activity. *Health Psychology* 34(12): 1145–1153. Available from: http://psycnet.apa.org/journals/hea/34/12/1145/%5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/hea0000223.

Edwards EA, Lumsden J, Rivas C, et al. (2016) Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps. *BMJ open* 6(10): e012447.

Eisenhauer CM, Hageman PA, Rowland S, et al. (2017) Acceptability of mHealth technology for self-monitoring eating and activity among rural men. *Public Health Nursing* 34(2): 138–146.

Ekkekakis P (2003) Pleasure and displeasure from the body: Perspectives from exercise. *Cognition & Emotion* 17(June 2015): 213–239.

Ekkekakis P (2017) People have feelings! Exercise psychology in paradigmatic transition. *Current Opinion in Psychology* 16: 84–88.

Ekkekakis P and Dafermos M (2012) Exercise is a many-splendored thing, but for some it does not feel so splendid: staging a resurgence of hedonistic ideas in the quest to understand exercise behavior. In: Acevedo EO (ed.), *The Oxford Handbook of Exercise Psychology*, Oxford University Press.

Ekkekakis P and Lind E (2006) Exercise does not feel the same when you are overweight: the impact of self-selected and imposed intensity on affect and exertion. *International Journal of Obesity* 30(4): 652–660.

Available from: http://www.nature.com/doifinder/10.1038/sj.ijo.0803052.

Ekkekakis P, Hall EE, VanLanduyt LM, et al. (2000) Walking in (affective) circles: Can short walks enhance affect? *Journal of Behavioral Medicine* 23(3): 245–275.

Ekkekakis P, Hall EE and Petruzzello SJ (2005) Variation and homogeneity in affective responses to physical activity of varying intensities: An alternative perspective on dose-response based on evolutionary considerations. *Journal of Sports Sciences* 23(5): 477–500.

Ekkekakis P, Hargreaves EA and Parfitt G (2013) Invited Guest Editorial: Envisioning the next fifty years of research on the exercise-affect relationship. *Psychology of Sport and Exercise* 14(5): 751–758.

Ellingson LD, Lansing JE, DeShaw KJ, et al. (2019) Evaluating Motivational Interviewing and habit formation to enhance the effect of activity trackers on healthy adults' activity levels: randomized intervention. *JMIR mHealth and uHealth* 7(2): e10988.

Emerson JA and Williams DM (2015) The multifaceted relationship between physical activity and affect. *Social and Personality Psychology Compass* 9(8): 419–433.

Epton T, Currie S and Armitage CJ (2017) Unique effects of setting goals on behavior change: systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology* 85(12): 1182–1198.

Erriquez E and Grasso F (2008) Generation of personalised advisory messages: An ontology based approach. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*: 437–442.

Ess C and AoIR ethics working committee (2011) Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee. *Readings in Virtual Research Ethics*: 27–44.

Eysenbach G (2001) What is e-health? *Journal of Medical Internet Research* 3(2): 1–5.

Fahim M, Idris M, Ali R, et al. (2014) ATHENA: a personalized platform to promote an active lifestyle and wellbeing based on physical, mental and social health primitives. *Sensors (Basel, Switzerland)* 14(5): 9313–9329. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=24859031&site=ehost-live.

Fallaize R, Zenun Franco R, Pasang J, et al. (2019) Popular nutrition-related mobile apps: An agreement assessment against a UK reference method. *JMIR mHealth and uHealth* 7(2): e9838.

Fanning J, Mullen SP and McAuley E (2012) Increasing physical activity with mobile devices: a meta-analysis. *Journal of medical Internet research* 14(6): e161. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3514847&tool=pmcentrez&rendertype=abstract (accessed 10 July 2014).

Fanning J, Roberts S, Hillman CH, et al. (2017) A smartphone “app”-delivered randomized factorial trial targeting physical activity in adults. *Journal of Behavioral Medicine*, Springer US 40(5): 712–729.

Fatima I, Halder S, Saleem MA, et al. (2015) Smart CDSS: integration of Social Media and Interaction Engine (SMIE) in healthcare for chronic disease patients. *Multimedia Tools and Applications* 74(14): 5109–5129.

Federal Trade Commission (n.d.) Mobile Health Apps Interactive Tool. Available from: https://www.ftc.gov/tips-advice/business-center/guidance/mobile-health-apps-interactive-tool (accessed 26 July 2019).

Fernández FG, Solà MP, Ángel M, et al. (2013) Using smartphone bases biodevices for analyzing physiological, psychological and behavioral user’s habits. In: *BIODEVICES 2013 - International conference on biomedical electronics and devices*, pp. 243–248.

Finlay L (2002) Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. *Qualitative Research* 2(2): 209–230.

Fishbein M and Ajzen I (2010) *Predicting and changing behavior: The reasoned action approach*. New York: Taylor & Francis.

Fitbit (n.d.) Fitbit App. Available from: www.fitbit.com/uk/app.

Flick U and Tiidenberg K (2018) Ethics in digital research. *The SAGE Handbook of Qualitative Data Collection* (May): 466–479.

Focht BC (2009) Brief walks in outdoor and laboratory environments. *Research Quarterly for Exercise and Sport* 80(3): 611–620. Available from: http://www.tandfonline.com/doi/abs/10.1080/02701367.2009.10599600.

Focht BC (2013) Affective responses to 10-minute and 30-minute walks in sedentary, overweight women: Relationships with theory-based correlates of walking for exercise. *Psychology of Sport and Exercise*, Elsevier Ltd 14(5): 759–766. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1469029213000344.

Fogg BJ (2003) *Persuasive technology: Using computers to change what we think and do*. San Francisco: Morgan Kaufmann Publishers.

Ford ES, Merritt RK, Heath GW, et al. (1991) Physical activity behaviors in lower and higher socioeconomic status populations. *American Journal of Epidemiology* 133(12): 1246–1256.

Forsblom N (2015) Were you aware of all these sensors in your smartphone? *Adtile*. Available from: https://blog.adtile.me/2015/11/12/were-you-aware-of-all-these-sensors-in-your-smartphone/.

Forster AS, Buykx P, Martin N, et al. (2017) Using affective judgement to increase physical activity in British adults. *Health Promotion International*: 1–9. Available from: http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf.

Fougerouse P, Yasini M, Marchand G, et al. (2017) A cross-sectional study of prominent US mobile health applications: evaluating the current landscape. *AMIA Annu Syp Proc*: 715–723.

Fox KR (2007) The influence of physical activity on mental well-being. *Public Health Nutrition*, Cambridge University Press 2(3a): 411–418. Available from: http://journals.cambridge.org/abstract_S1368980099000567 (accessed 18 November 2014).

Fox S and Duggan M (2012) *Mobile Health 2012*. *PewResearch Centre's Internet & American Life Project Washington, D.C.* Available from: http://www.pewinternet.org/~/media//Files/Reports/2012/PIP_MobileHealth2012_FINAL.pdf.

Fox S and Duggan M (2013) Tracking for health. *Pew Internet* (October): 1–40. Available from: http://www.pewinternet.org/Reports/2013/Tracking-for-Health.aspx.

Free C, Phillips G, Galli L, et al. (2013) The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review. *PLoS Medicine* 10(1): e1001362.

Freeman B, Chau J and Mihrshahi S (2017) Why the public health sector couldn't create Pokémon Go. *Public Health Research & Practice* 27(3): 1–3. Available from: http://www.phrp.com.au/?p=36697.

Freis A, Freundl-Schütt T, Wallwiener L-M, et al. (2018) Plausibility of menstrual cycle apps claiming to support conception. *Frontiers in Public Health* 6(April): 1–9. Available from: http://journal.frontiersin.org/article/10.3389/fpubh.2018.00098/full.

Fujiki Y, Kazakos K, Puri C, et al. (2008) NEAT-o-Games: Blending physical activity and fun in the daily routine. *ACM Comput. Entertain.* 6(1): 21. Available from: http://doi.acm.org/10.1145/1371216.1371224.

Garcia D and Archer T (2014) Positive affect and age as predictors of exercise compliance. *PeerJ.* 2(2167-8359 (Electronic)): e694.

Gardner B and Rebar AL (2019) *Habit Formation and Behavior Change*.

Psychology.

Gardner B, Smith L, Lorencatto F, et al. (2015) How to reduce sitting time? A review of behaviour change strategies used in sedentary behaviour reduction interventions among adults. *Health psychology review*, Taylor & Francis 7199(October): 1–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26315814.

Garnett C, Crane D, Brown J, et al. (2018) Reported theory use by digital interventions for hazardous and harmful alcohol consumption, and association with effectiveness: Meta-regression. *Journal of Medical Internet Research* 20(2).

Gasser R, Brodbeck D, Degen M, et al. (2006) Persuasiveness of a mobile lifestyle coaching application using social facilitation. In: *Lecture Notes in Computer Science*, pp. 1–6.

Gauvin L and Brawley LR (1993) Alternative psychological models and methodologies for the study of exercise and affect. In: Seraganian P (ed.), *Exercise psychology: The influence of physical exercise on pyschological processes*, New York: John Wiley & Sons, pp. 146–171.

Ghanvatkar S, Kankanhalli A and Rajan V (2019) User models for personalized physical activity interventions: Scoping review. *Journal of Medical Internet Research* 21(1).

Gill DL, Hammond CC, Reifsteck EJ, et al. (2013) Physical activity and quality of life. *Journal of preventive medicine and public health = Yebang Ŭihakhoe chi* 46 Suppl 1: S28-34. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3567315&tool=pmcentrez&rendertype=abstract.

Glanz K and Bishop DB (2010) The role of behavioral science theory in development and implementation of Public Health interventions. *Annual Review of Public Health* 31(1): 399–418. Available from: http://www.annualreviews.org/doi/10.1146/annurev.publhealth.012809.103604.

Glasgow R, Vogt T and Boles S (1999) Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health* 89(9): 1322–1327. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1508772/.

Google (2019) Google Play Terms of Service. Available from: https://play.google.com/intl/en_uk/about/play-terms/index.html (accessed 29 July 2019).

Google (n.d.) Comment posting policy. Available from: https://play.google.com/about/comment-posting-policy.html?hl=en-GB (accessed 29 July 2019).

Goto M, Takedani H, Haga N, et al. (2014) Self-monitoring has potential for home exercise programmes in patients with haemophilia. *Haemophilia: The Official Journal Of The World Federation Of Hemophilia* 20(2): e121–e127. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=24418413&site=ehost-live.

Grady HM (2000) Web site design: A case study in usability testing using paper prototypes. In: *Proceedings of the 18th annual ACM international conference on computer documentation: technology & teamwork*, pp. 39–45.

Grant MJ and Booth A (2009) A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal* 26(2): 91–108.

Gray DE (2018) *Doing research in the real world*. 4th ed. London: Sage Publications Ltd.

Greaves C, Sheppard KE, Abraham C, et al. (2011) Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. *BMC Public Health* 11(119). Available from: http://www.biomedcentral.com/content/pdf/1471-2458-11-119.pdf (accessed 9 October 2014).

Greene JC, Caracelli VJ and Graham WF (1989) Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis* 11(3): 255–274.

Griffiths C, Harnack L and Pereira MA (2018) Assessment of the accuracy of nutrient calculations of five popular nutrition tracking applications. *Public Health Nutrition* 21(8): 1495–1502.

Griffiths KM and Christensen H (2000) Information in practice depression: cross sectional survey. *Bmj* 321: 1511–1515.

Guba EG (1990) The alternative paradigm dialog. In: *The paradigm dialog*, London: SAGE Publications, pp. 17–30. Available from: http://www.jstor.org/stable/3340973.

Guba EG and Lincoln YS (1994) Competing paradigms in qualitative research. In: Denzin, Norman K and Lincoln, Yvonna S (eds), *Handbook of Qualitative Research*, SAGE Publications, pp. 105–117.

Guérin E, Fortier MS and Sweet SN (2013) An experience sampling study of physical activity and positive affect: investigating the role of situational motivation and perceived intensity across time. *Health Psychology Research* 1(e21).

Guess N (2012) A qualitative investigation of attitudes towards aerobic and resistance exercise amongst overweight and obese individuals. *BMC Research Notes* 5(191). Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3490848&tool=pmcentrez&rendertype=abstract (accessed 15 October 2014).

Guyatt G, Cairns J, Churchill D, et al. (1992) Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA: the Journal of the American Medical Association* 268(17): 2420–5.

Guzman E and Maalej W (2014) How do users like this feature? A fine grained sentiment analysis of App reviews. In: *2014 IEEE 22nd*

International Requirements Engineering Conference, Karlskrona, Sweden, IEEE, pp. 153–162.

Hall PA and Fong GT (2007) Temporal self-regulation theory: A model for individual health behavior. *Health Psychology Review* 1(1): 6–52. Available from: http://www.tandfonline.com/doi/abs/10.1080/17437190701492437 (accessed 10 October 2014).

Hallal PC, Andersen LB, Bull FC, et al. (2012) Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet*, Elsevier Ltd 380(9838): 247–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22818937 (accessed 16 September 2013).

Hansen CJ, Stevens LC and Coast JR (2001) Exercise duration and mood state: How much is enough to feel better? *Health Psychology* 20(4): 267–275.

Hardeman W, Lamming L, Kellar I, et al. (2014) Implementation of a nurse-led behaviour change intervention to support medication taking in type 2 diabetes: beyond hypothesised active ingredients (SAMS Consultation Study). *Implementation science: IS* 9(1): 70. Available from: http://www.implementationscience.com/content/9/1/70.

Hardeman W, Houghton J, Lane K, et al. (2019) A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity* 16(1).

Hardy CJ and Rejeski WJ (1989) Not what, but how one feels: The measurement of affect during exercise. *Journal of Sport & Exercise Psychology* 11: 304–317.

Hardy S and Grogan S (2009) Preventing disability through exercise: investigating older adults' influences and motivations to engage in physical activity. *Journal of Health Psychology* 14(7): 1036–1046. Available from:

http://hpq.sagepub.com/content/14/7/1036.abstract%5Cnhttp://hpq.sagepub.com/cgi/doi/10.1177/1359105309342298.

Harkin B, Webb T, Chang B, et al. (2016) Does monitoring goal progress promote goal attainment? A meta-analysis of the experimental evidence. *Psychological Bulletin* 142(2): 198–229.

Harland J, White M, Drinkwater C, et al. (1999) The Newcastle exercise project: a randomised controlled trial of methods to promote physical activity in primary care. *BMJ (Clinical research ed.)* 319(7213): 828–32. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=314213&tool=pmcentrez&rendertype=abstract.

Harris BS, Melton B, Bland H, et al. (2018) Enhancing psychosocial constructs associated with technology-based physical activity: a randomized trial among african american women. *American Journal of Health Education* 49(2): 74–85.

Hassandra M, Lintunen T, Hagger MS, et al. (2017) An mHealth app for supporting quitters to manage cigarette cravings with short bouts of physical activity: a randomized pilot feasibility and acceptability study. *JMIR mHealth and uHealth* 5(5): e74. Available from: http://mhealth.jmir.org/2017/5/e74/.

Hayes AF and Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1): 77–89. Available from: http://www.tandfonline.com/doi/abs/10.1080/19312450709336664.

HealthTap (2010). Available from: https://www.healthtap.com/ (accessed 29 July 2019).

Hearn L, Miller M and Lester L (2014) Reaching perinatal women online: the Healthy You, Healthy Baby website and app. *Journal Of Obesity* 2014: 573928. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=24

872891&site=ehost-live.

Hekler EB, Buman MP, Grieco L, et al. (2015) Validation of physical activity tracking via Android smartphones compared to ActiGraph accelerometer: laboratory-based and free-living validation studies. *JMIR mHealth and uHealth* 3(2): e36. Available from: http://mhealth.jmir.org/2015/2/e36/.

Hendershot CS, Witkiewitz K, George WH, et al. (2011) Relapse prevention for addictive behaviors. *Substance Abuse: Treatment, Prevention, and Policy* 6(1): 1–17.

Higgins JPT and Green S (eds) (2011) *Cochrane Handbook for Systematic Reviews of Interventions v5.1.0*. The Cochrane Collaboration. Available from: www.handbook.cochrane.org.

Higgins JPT, Savović J, Page MJ, et al. (2016) Revised Cochrane risk of bias tool for randomized trials (RoB 2.0). *Cochrane Methods* (10 (Suppl 1)): 52.

Hoddinott P, Allan K, Avenell A, et al. (2010) Group interventions to improve health outcomes: A framework for their design and delivery. *BMC Public Health*, BioMed Central Ltd 10(1): 800. Available from: http://www.biomedcentral.com/1471-2458/10/800.

Hoffmann TC, Glasziou PP, Boutron I, et al. (2014) Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Online)* 348(March): 1–12. Available from: http://dx.doi.org/doi:10.1136/bmj.g1687.

Hogan CL and Carstensen LL (2013) Exercise holds immediate benefits for affect and cognition in younger and older adults. *Psychol Aging* 28(2): 587–594.

Hoj TH, Covey EL, Jones AC, et al. (2017) How do apps work? An analysis of physical activity app users’ perceptions of behavior change mechanisms. *JMIR mHealth and uHealth* 5(8): e114. Available from: http://mhealth.jmir.org/2017/8/e114/.

Holst A (2019) Global smartphone sales to end users from 1st quarter 2009 to 2nd quarter 2018, by operating system (in million units). *STATISTA*. Available from: https://www.statista.com/statistics/266219/global-smartphone-sales-since-1st-quarter-2009-by-operating-system/ (accessed 29 July 2019).

Howlett N, Trivedi D, Troop NA, et al. (2018) Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? A systematic review and meta-analysis. *Translational Behavioral Medicine* (May): 1–11. Available from: https://academic.oup.com/tbm/advance-article/doi/10.1093/tbm/iby010/4913688.

Hu N, Pavlou PA and Zhang J (2006) Can online reviews reveal a product's true quality?: 324–330.

Idris M, Hussain S, Ahmad M, et al. (2015) Big Data service engine (BISE): Integration of big data technologies for human centric wellness data. *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015*, IEEE: 244–248.

iMedical Apps (n.d.) *MedPage Today*. Available from: www.imedicalapps.com (accessed 29 July 2019).

Iqbal M (2019) App download and usage statistics (2019). *Business of Apps*. Available from: https://www.businessofapps.com/data/app-statistics/#6 (accessed 2 August 2019).

IQVIA (n.d.) AppScript. Available from: https://www.appscript.net/ (accessed 29 July 2019).

Jake-Schoffman DE, Silfee VJ, Waring ME, et al. (2017) Methods for evaluating the content, usability, and efficacy of commercial mobile health apps. *JMIR mHealth and uHealth* 5(12): e190. Available from: http://mhealth.jmir.org/2017/12/e190/.

Jee H (2017) Review of researches on smartphone applications for physical activity promotion in healthy adults. *Journal of Exercise Rehabilitation* 13(1): 3–11. Available from: http://e-jer.org/journal/view.php?number=2013600338.

Jekauc D and Brand R (2017) Editorial: How do emotions and feelings regulate physical activity? *Frontiers in Psychology* 8(JUL): 1–3.

Jeon E, Park HA, Min YH, et al. (2014) Analysis of the information quality of Korean obesity-management smartphone applications. *Healthcare Informatics Research* 20(1): 23–29.

Jin M and Kim J (2015) Development and evaluation of an evaluation tool for healthcare smartphone applications. *Telemedicine and e-Health* 21(10): 831–837. Available from: http://online.liebertpub.com/doi/10.1089/tmj.2014.0151.

Johnston M, Carey RN, Connell LE, et al. (2018) *Linking behaviour change techniques and mechanisms of action: Triangulation of findings from literature synthesis and expert consensus*. Available from: https://doi.org/10.31234/osf.io/ur6kz.

Joint Health Surveys Unit (2009) *Health Survey for England 2008. Volume 1 Physical activity and fitness. The Stationary Office.* Available from: http://www.hscic.gov.uk/pubs/hse08physicalactivity.

Kanjo E, Younis EMG and Sherkat N (2018) Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach. *Information Fusion*, Elsevier 40: 18–31.

Kanning M, Ebner-Priemer U and Schlicht W (2015) Using activity triggered e-diaries to reveal the associations between physical activity and affective states in older adult's daily living. *International Journal of Behavioral Nutrition and Physical Activity* 12(1): 111. Available from: http://www.ijbnpa.org/content/12/1/111.

Kaplan WA (2006) Can the ubiquitous power of mobile phones be used to improve health outcomes in developing countries? *Globalization and*

Health 2: 1–14.

Karduck J and Chapman-Novakofski K (2018) Results of the clinician apps survey, how clinicians working with patients with diabetes and obesity use mobile health apps. *Journal of Nutrition Education and Behavior*, Elsevier Inc. 50(1): 62-69.e1. Available from: https://doi.org/10.1016/j.jneb.2017.06.004.

Kesiraju L and Vogels T (2017) Health & Fitness app users are going the distance with record-high engagement. *FlurryMobile*. Available from: https://lmphil.tumblr.com/post/186188441928/health-fitness-app-users-are-going-the-distance (accessed 2 August 2019).

Khalid H, Shihab E, Nagappan M, et al. (2014) What do mobile app users complain about? A study on free iOS apps. *IEEE Software* 32(3): 70–77. Available from: http://ieeexplore.ieee.org/document/6762802/.

Kim KK, Logan HC, Young E, et al. (2015) Youth-centered design and usage results of the iN Touch mobile self-management program for overweight/obesity. *Personal and Ubiquitous Computing* 19(1): 59–68.

Kim YS, Park YS, Allegrante JP, et al. (2012) Relationship between physical activity and general mental health. *Preventive Medicine*, Elsevier Inc. 55(5): 458–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22981733 (accessed 22 May 2013).

King AC, Ahn DK, Oliveira BM, et al. (2008) Promoting physical activity through hand-held computer technology. *American Journal of Preventive Medicine* 34(2): 138–142. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0749379707006459.

Klein M, Mogles N and van Wissen A (2014) Intelligent mobile support for therapy adherence and behavior change. *Journal of Biomedical Informatics*, Elsevier Inc. 51: 137–151. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1532046414001221.

Klenk S, Reifegerste D and Renatus R (2017) Gender differences in gratifications from fitness app use and implications for health interventions. *Mobile Media and Communication* 5(2): 178–193.

Klock ACT and Gasparini I (2015) A usability evaluation of fitness-tracking apps for initial users. *Communications in Computer and Information Science* 529: 457–462. Available from: http://link.springer.com/10.1007/978-3-319-40548-3.

Knight E, Stuckey MI, Prapavessis H, et al. (2015) Public Health Guidelines for physical activity: Is there an app for that? A review of Android and Apple app stores. *JMIR Mhealth and Uhealth* 643(32): 33–43. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4456485/?report=printable [10/6/%5Cnhttp://mhealth.jmir.org.

Knox ECL, Esliger DW, Biddle SJH, et al. (2013) Lack of knowledge of physical activity guidelines: can physical activity promotion campaigns do better ? *BMJ Open* 3: 1–7.

Knox ECL, Webb OJ, Esliger DW, et al. (2013) Using threshold messages to promote physical activity: Implications for public perceptions of health effects. *The European Journal of Public Health* 1–5.

Knuth AG and Hallal PC (2009) Temporal trends in physical activity: A systematic review. *Journal of Physical Activity and Health* 6(5): 548–559. Available from: http://journals.humankinetics.com/doi/10.1123/jpah.6.5.548.

König LM, Sproesser G, Schupp HT, et al. (2018) Describing the process of adopting nutrition and fitness apps: Behavior stage model approach. *Journal of Medical Internet Research* 20(3): 1–15.

Krippendorff K (2011) Computing Krippendorff’s alpha-reliability. *Departmental Papers (ASC)*: 12. Available from: http://repository.upenn.edu/asc_papers.

Kwan B and Bryan A (2010) In-task and post-task affective response to exercise: translating exercise intentions into behaviour. *British Journal of*

Health Psychology 15(Pt 1): 115–131.

Kwasnicka D, Dombrowski SU, White M, et al. (2016) Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. *Health Psychology Review* 10(3): 277–296.

Lamattina S (2016) How to find app store keywords that people actually search for. *App Tool Tester*. Available from: https://apptooltester.com/app-store-keywords/ (accessed 29 July 2019).

Lamming L, Mason D, Morton K, et al. (2017) What do we know about brief interventions for physical activity. *Preventive Medicine*, Elsevier Inc. 44(0): 152–163. Available from: http://dx.doi.org/10.1016/j.ypmed.2017.02.017.

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Lane AM, Jackson A and Terry PC (2005) Preferred modality influences on exercise-induced mood changes. *Journal of Sports Science and Medicine* 4(2): 195–200.

Lathia N, Pejovic V, Rachuri KK, et al. (2013) Smartphones for large-scale behavior change interventions. *IEEE Pervasive Computing* 12: 66–73.

Lathia N, Sandstrom GM, Mascolo C, et al. (2017) Happier people live more active lives: Using smartphones to link happiness and physical activity. *PLoS ONE* 12(1): 1–13.

Laverie DA (1998) Motivations for ongoing participation in a fitness activity. *Leisure Sciences* 20(4): 277–302. Available from: http://www.tandfonline.com/doi/abs/10.1080/01490409809512287.

Lawton J, Ahmad N, Hanna L, et al. (2006) ‘I can’t do any serious exercise’: Barriers to physical activity amongst people of Pakistani and Indian origin with Type 2 diabetes. *Health Education Research* 21(1): 43–54.

Lee I-M, Shiroma EJ, Lobelo F, et al. (2012) Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*, Elsevier Ltd 380(9838): 219–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22818936 (accessed 9 February 2013).

Lewis-Beck MS, Bryman A and Liao TF (2004) The SAGE encyclopedia of qualitative research methods Vol.3.: 168–170.

Lewis BA, Napolitano MA, Buman MP, et al. (2017) Future directions in physical activity intervention research: expanding our focus to sedentary behaviors, technology, and dissemination. *Journal of Behavioral Medicine*, Springer US 40(1): 112–126.

Liao Y (2015) Understanding the dynamic relationships between physical activity and affective states using real-time data capture techniques. University of Southern California.

Liao Y, Shonkoff ET and Dunton GF (2015) The acute relationships between affect, physical feeling states, and physical activity in daily life: A review of current evidence. *Frontiers in Psychology* 6(1975): 1–7.

Lincoln YS and Guba EG (1985) *Naturalistic Inquiry*. 1st ed. London: Sage Publications Ltd.

Lox CL, Jackson S, Tuholski SW, et al. (2000) Revisiting the measurement of exercise-induced feeling states: The Physical Activity Affect Scale. *Measurement in Physical Education and Exercise Science* 4(2): 79–95.

Loy JS, Ali EE and Yap KY-L (2016) Quality assessment of medical apps that target medication-related problems. *Journal of Managed Care & Specialty Pharmacy* 22(10): 1124–1140.

Lyons EJ, Lewis ZH, Mayrsohn BG, et al. (2014) Behavior change techniques implemented in electronic lifestyle activity monitors: A systematic content analysis. *Journal of Medical Internet Research* 16(8): e192.

Lyons EJ, Swartz MC, Lewis ZH, et al. (2017) Feasibility and acceptability of a wearable technology physical activity intervention with telephone counseling for mid-aged and older adults: A randomized controlled pilot trial. *JMIR mHealth and uHealth* 5(3): e28.

Lyzwinski LN (2014) A systematic review and meta-analysis of mobile devices and weight loss with an intervention content analysis. *Journal of Personalized Medicine* 4(3): 311–385.

Maes S and Karoly P (2005) Self-regulation assessment and intervention in physical health and illness: A review. *Applied Psychology* 54(2): 267–299.

Magnan RE, Kwan BM and Bryan AD (2013) Effects of current physical activity on affective response to exercise: Physical and social–cognitive mechanisms. *Psychology & Health* 28(4): 418–433. Available from: http://www.tandfonline.com/doi/abs/10.1080/08870446.2012.733704.

Maitland J, Sherwood S, Barkhuus L, et al. (2007) Increasing the awareness of daily activity levels with pervasive computing. *2006 Pervasive Health Conference and Workshops, PervasiveHealth*: 1–9.

Maltseva K and Lutz C (2018) A quantum of self: A study of self-quantification and self-disclosure. *Computers in Human Behavior*, Elsevier Ltd 81: 102–114.

Marceau LD, Link C, Jamison RN, et al. (2007) Electronic diaries as a tool to improve pain management: Is there any evidence? *Pain Medicine* 8(SUPPL.3).

Markham A (2012) AOIR Guidelines: Ethical decision making and internet research ethics. Appendix 1. Available from: https://aoir.org/wp-content/uploads/2017/01/aoir_ethics_graphic_2016.pdf.

Martin SS, Feldman DI, Blumenthal RS, et al. (2015) mActive: A randomized clinical trial of an automated mHealth intervention for physical activity promotion. *Journal of the American Heart Association* 4(11): 1–9.

Martínez-García MDM, Ruiz-Cárdenas JD and Rabinovich RA (2017) Effectiveness of smartphone devices in promoting physical activity and exercise in patients with chronic obstructive pulmonary disease: A systematic review. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 14(5): 543–551. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L618243193%0Ahttp://dx.doi.org/10.1080/15412555.2017.1358257.

Martinez-Nicolas A, Muntaner-Mas A and Ortega FB (2017) Runkeeper: A complete app for monitoring outdoor sports. *British Journal of Sports Medicine* 51(21): 1560–1561.

Martinez-Perez B, de la Torre-Diez I and Lopez-Coronado M (2015) Experiences and results of applying tools for assessing the quality of a mHealth app named Heartkeeper. *Journal of Medical Systems* 39(11).

Martínez-Pérez B, De La Torre-Díez I, Candelas-Plasencia S, et al. (2013) Development and evaluation of tools for measuring the quality of experience (QoE) in mHealth applications. *Journal of Medical Systems* 37(5).

Martinez N, Kilpatrick MW, Salomon K, et al. (2015) Affective and enjoyment responses to high-intensity interval training in overweight-to-obese and insufficiently active adults. *Journal of Sport & Exercise Psychology* 37: 138–149.

Mateo GF, Granado-Font E, Ferre-Grau C, et al. (2015) Mobile phone apps to promote weight loss and increase physical activity: A systematic review and meta-analysis. *Journal of Medical Internet Research* 17(11): 1–14.

Mays N (2002) Qualitative research in health care: Assessing quality in qualitative research. *BMJ* 320(7226): 50–52.

McAuley E, Jerome GJ, Elavsky S, et al. (2003) Predicting long-term maintenance of physical activity in older adults. *Preventive Medicine* 37(2): 110–118. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0091743503000896.

McCaslin ML (2008) Pragmatism. In: Given LM (ed.), *The Sage encyclopedia of qualitative research methods*, Thousand Oaks: SAGE Publications, pp. 672–675.

McEwan D, Beauchamp MR, Kouvousis C, et al. (2018) Examining the active ingredients of physical activity interventions underpinned by theory versus no stated theory: a meta-analysis. *Health Psychology Review*, Taylor & Francis 0(0): 1–17. Available from: https://www.tandfonline.com/doi/full/10.1080/17437199.2018.1547120.

Mcguire A, Seib C and Anderson D (2016) Factors predicting barriers to exercise in midlife Australian women. *Maturitas*, Elsevier Ireland Ltd 87: 61–66.

McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochemia Medica* 22(3): 276–282. Available from: http://www.biochemia-medica.com/node/501.

McIlroy S, Ali N, Khalid H, et al. (2016) Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empirical Software Engineering*, Empirical Software Engineering 21(3): 1067–1106. Available from: http://dx.doi.org/10.1007/s10664-015-9375-7.

McLeroy KR, Bibeau D, Steckler A, et al. (1988) An ecological perspective on health promotion programs. *Health Education Quarterly* 15(4): 351–377.

McMillan B, Hickey E, Patel MG, et al. (2015) Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance. *Patient Education and Counseling*, Elsevier Ireland Ltd 99(3): 429–435. Available from: http://dx.doi.org/10.1016/j.pec.2015.10.023.

McNiel P and McArthur EC (2016) Evaluating health mobile apps: Information literacy in undergraduate and graduate nursing courses. *Journal Nurs Education* 55(8): 480.

Medicines & Healthcare Products Regulatory Agency (2018) Guidance: Medical device stand-alone software including apps.: 1–39. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/717865/Software_flow_chart_Ed_1-05.pdf.

Melby C, Scholl C, Edwards G, et al. (1993) Effect of acute resistance exercise on postexercise energy expenditure and resting metabolic rate. *Journal of Applied Physiology* 75(4): 1847–1853. Available from: http://www.physiology.org/doi/10.1152/jappl.1993.75.4.1847.

Melton BF, Buman MP, Vogel RL, et al. (2016) Wearable devices to improve physical activity and sleep: A randomized controlled trial of college-aged African American women. *Journal of Black Studies* 47(6): 610–625.

Mendiola MF, Kalnicki M and Lindenauer S (2015) Valuable features in mobile health apps for patients and consumers: content analysis of apps and user ratings. *JMIR mHealth and uHealth* 3(2): e40. Available from: http://mhealth.jmir.org/2015/2/e40/.

Mertens DM, Bledsoe KL, Sullivan M, et al. (2010) Utilization of mixed methods for transformative purposes. 2nd Editio. In: Tashakkori A and Teddlie C (eds), *SAGE Handbook of Mixed Methods in Social and Behavioural Research*, London: Sage Publications Ltd.

Michie S and Prestwich A (2010) Are interventions theory-based? Development of a theory coding scheme. *Health Psychology: official journal of the Division of Health Psychology, American Psychological Association* 29(1): 1–8.

Michie S, Abraham C, Whittington C, et al. (2009) Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychology: official journal of the Division of Health Psychology, American Psychological Association* 28(6): 690–701. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19916637 (accessed 21 July 2014).

Michie S, van Stralen MM and West R (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science: IS*, BioMed Central Ltd 6(1): 42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096582&tool=pmcentrez&rendertype=abstract (accessed 9 July 2014).

Michie S, Richardson M, Johnston M, et al. (2013) The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine* 46(1): 81–95. Available from: http://link.springer.com/10.1007/s12160-013-9486-6.

Michie S, West R, Campbell R, et al. (2014) *ABC of behaviour change theories: an essential resource for researchers, policy makers and practitioners*. Silverback Publishing. Available from: https://www.waterstones.com/book/abc-of-behaviour-change-theories/susan-michie/robert-west/9781912141012.

Michie S, Atkins L and West R (2015) *The Behaviour Change Wheel - A Guide to Designing Interventions*. Silverback Publishing.

Michie S, Yardley L, West R, et al. (2017) Developing and evaluating digital interventions to promote behavior change in health and health care: Recommendations resulting from an international workshop. *Journal of Medical Internet Research* 19(6).

Michie S, West R, Sheals K, et al. (2018) Evaluating the effectiveness of behavior change techniques in health-related behavior: A scoping review of methods used. *Translational Behavioral Medicine* 8(2): 212–224.

Middelweerd A, Mollee JS, van der Wal CN, et al. (2014) Apps to promote physical activity among adults: a review and content analysis. *The International Journal of Behavioral Nutrition and Physical Activity* 11(1): 97. Available from: http://www.ijbnpa.org/content/11/1/97 (accessed 29 July 2014).

Middelweerd A, van der Laan DM, van Stralen MM, et al. (2015) What features do Dutch university students prefer in a smartphone application for promotion of physical activity? A qualitative approach. *International Journal of Behavioral Nutrition and Physical Activity* 12(1): 31. Available from: http://www.ijbnpa.org/content/12/1/31.

Miller AS, Cafazzo JA and Seto E (2014) A game plan: Gamification design principles in mHealth applications for chronic disease management. *Health Informatics Journal* 22(2): 184–193.

Miller S, Ainsworth B, Yardley L, et al. (2018) A framework for Analysing and Measuring Usage and Engagement Data (AMUsED) in digital interventions: Viewpoint. *Journal of Medical Internet Research* 21: 1–13.

Milward J, Khadjesari Z, Fincham-Campbell S, et al. (2016) User preferences for content, features, and style for an app to reduce harmful drinking in young adults: Analysis of user feedback in app stores and focus group interviews. *JMIR Mhealth And Uhealth* 4(2): e47–e47. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=27220371&site=ehost-live.

Mingers J (2006) Chapter 2: Philosophical Foundations: Critical Realism. *Systems Thinking, Critical Realism and Philosophy*. Available from: https://www.taylorfrancis.com/books/9781315774503.

Moglia ML, Nguyen H V., Chyjek K, et al. (2016) Evaluation of smartphone menstrual cycle tracking applications using an adapted applications scoring system. *Obstetrics and Gynecology* 127(6): 1153–1160.

Moher D, Liberati A, Tetzlaff J, et al. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 339(jul21 1): b2535–b2535. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.b2535 (accessed 21 May 2013).

Monroe CM, Thompson DL, Bassett DR, et al. (2015) Usability of mobile phones in physical activity–related research: A systematic review.

American Journal of Health Education 46(4): 196–206. Available from: http://www.tandfonline.com/doi/full/10.1080/19325037.2015.1044141.

Montori VM, Wang YG, Alonso-Coello P, et al. (2006) Systematic evaluation of the quality of randomized controlled trials in diabetes. *Diabetes Care* 29(8): 1833–1838.

Moore GF, Audrey S, Barker M, et al. (2015) Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 350(mar19 6): h1258–h1258. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.h1258.

Morandi A and Serafin R (2007) A personalized motivation strategy for physical activity promotion in diabetic subjects. In: *Proceedings of the 2nd Workshop on Personalisation for E-Health*, pp. 51–55.

Mulrow CD (1994) Systematic Reviews: Rationale for systematic reviews. *Bmj.Com* 309: 597–599. Available from: https://www.bmj.com/content/309/6954/597?variant=extract.

Muntaner A, Vidal-Conti J and Palou P (2015) Increasing physical activity through mobile device interventions: A systematic review. *Health Informatics Journal*. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25649783.

Murray E, Hekler EB, Andersson G, et al. (2016) Evaluating digital health interventions: Key questions and approaches. *American Journal of Preventive Medicine*, Elsevier 51(5): 843–851. Available from: http://dx.doi.org/10.1016/j.amepre.2016.06.008.

Murray L (2006) *Sport, exercise and physical activity: public participation, barriers and attitudes*. *Scottish Executive*, Edinburgh:Scottish Executive Education Department.

Naslund JA, Marsch LA, McHugo GJ, et al. (2015) Emerging mHealth and eHealth interventions for serious mental illness: A review of the literature. *Journal of Mental Health* 24(5): 320–331.

National Health Service (2014) *Five Year Forward View*. Available from: https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf.

National Health Service (n.d.) NHS Apps Library. Available from: https://www.nhs.uk/apps-library/ (accessed 29 July 2019).

National Institute for Health Research (2017) National standards for public involvement in research. (June): 1–16. Available from: http://www.invo.org.uk/wp-content/uploads/2014/11/Draft-Values-principles-and-standards-framework-071114.pdf.

Nayebi M, Adams B and Ruhe G (2016) Release practices for mobile apps- what do users and developers think? In: *IEEE 23rd International Conference on Software ANalysis, Evolution, and Reengineering (SANER)*, Suita, Japan: IEEE. Available from: https://ieeexplore.ieee.org/abstract/document/7476674.

Neale C, Aspinall P, Roe J, et al. (2017) The aging urban brain: analyzing outdoor physical activity using the Emotiv Affectiv Suite in older people. *Journal of Urban Health* 94(6): 881.

Ng SW and Popkin BM (2012) Time use and physical activity: A shift away from movement across the globe. *Obesity Reviews* 13(8): 659–680.

NHS Digital (2017) *Health Survey for England 2016 Physical activity in adults*. Available from: https://files.digital.nhs.uk/publication/m/3/hse16-adult-phy-act.pdf.

NHS Digital (2018) *Digital Assessment Questionnaire V2.1*. Available from: https://www.england.nhs.uk/nhside.

NICE (2007) *Behaviour change: general approaches*. Available from: https://www.nice.org.uk/guidance/PH6.

NICE (2019) *Evidence standards framework for digital health technologies*. *NICE*. Available from: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/functional-classification-case-studies.pdf%0Ahttps://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evid.

Nielsen J (1993) *Usability engineering*. London: Academic Press Ltd.

Nielsen J (1994) Enhancing the explanatory power of usability heuristics. *Conference companion on Human factors in computing systems - CHI '94*: 210. Available from: http://portal.acm.org/citation.cfm?doid=259963.260333.

Nielsen J (2012) Usability 101: Introduction to Usability. *NN/g Nielsen Norman Group*. Available from: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ (accessed 2 August 2019).

Nieswiadomy RM (2012) Chapter 10: Qualitative research designs. *Foundations in nursing research*: 171–184.

Noar SM (2006) A 10-year retrospective of research in health mass media campaigns: Where do we go from here? *Journal of Health Communication* 11(1): 21–42.

Norman GJ and Velicer WF (2003) Developing an empirical typology for regular exercise. *Preventive Medicine* 37(6): 635–645.

Nouri R, R Niakan Kalhori S, Ghazisaeedi M, et al. (2018) Criteria for assessing the quality of mHealth apps: a systematic review. *Journal of the American Medical Informatics Association : JAMIA* 0(June): 1–10. Available from: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocy050/4996915%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/29788283.

O'Donoghue G, Perchoux C, Mensah K, et al. (2016) A systematic review of correlates of sedentary behaviour in adults aged 18–65 years: a socio-ecological approach. *BMC Public Health* 16(1): 163. Available from: http://www.biomedcentral.com/1471-2458/16/163.

O'Gorman K and MacIntosh R (2012) Research methods for business students. *The Global Management Series* (SEPTEMBER): 1–696.

O'Reilly GA and Spruijt-Metz D (2013) Current mHealth technologies for physical activity assessment and promotion. *American Journal of Preventive Medicine*, Elsevier 45(4): 501–507. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0749379713003942 (accessed 20 September 2013).

Ofcom (2015) *The Communications Market Report 2015*. Available from: https://www.ofcom.org.uk/research-and-data/multi-sector-research/cmr/cmr15.

Olander EK, Fletcher H, Williams S, et al. (2013) What are the most effective techniques in changing obese individuals' physical activity self-efficacy and behaviour: a systematic review and meta-analysis. *The International Journal of Behavioral Nutrition and Physical Activity* 10: 29. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3639155&tool=pmcentrez&rendertype=abstract.

Olsen L, Goolsby WA and McGinnis JM (2009) Institute of Medicine: Roundtable on Evidence-Based Medicine. In: *Leadership Commitments to Improve Value in Healthcare: Finding Common Ground Workshop Summary*, Washington: The National Academies Press. Available from: https://www.ncbi.nlm.nih.gov/books/NBK52843/#ch6.s1.

Onwuegbuzie A and Leech N (2005) On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology: Theory and Practice* 8(5): 375–387.

Oppezzo M and Schwartz DL (2014) Give your ideas some legs: The positive effect of walking on creative thinking. *Journal of Experimental Psychology: Learning Memory and Cognition* 40(4): 1142–1152.

ORCHA (n.d.). Available from: www.orcha.co.uk (accessed 24 July 2019).

Otto MW and Smits JAJ (2011) *Exercise for mood and anxiety: Proven* *strategies for overcoming depression and enhancing well being.* New York, NY: Oxford University Press. Available from: http://www.amazon.co.uk/Exercise-Mood-Anxiety-Strategies-Overcoming/dp/0199791007.

Ouellette JA and Wood W (1998) Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin* 124(1): 54–74.

Our Mobile Health (n.d.). Available from: http://www.ourmobilehealth.com/ (accessed 26 July 2019).

Pagoto S, Schneider K, Jojic M, et al. (2013) Evidence-based strategies in weight-loss mobile apps. *American Journal of Preventive Medicine*, Elsevier 45(5): 576–582. Available from: http://dx.doi.org/10.1016/j.amepre.2013.04.025.

Papon MTI, Ahmad I, Saquib N, et al. (2015) Non-invasive heart rate measuring smartphone applications using on-board cameras: A short survey. *Proceedings of 2015 International Conference on Networking Systems and Security, NSysS 2015*, IEEE: 1–6.

Parfitt G and Gledhill C (2004) The effect of choice of exercise mode on psychological responses. *Psychology of Sport and Exercise* 5(2): 111–117.

Parfitt G and Hughes S (2009) The exercise intensity–affect relationship: evidence and implications for exercise behavior. *Journal of Exercise Science & Fitness*, Elsevier (Singapore) Pte Ltd 7(2): S34–S41. Available from: http://www.sciencedirect.com/science/article/pii/S1728869X09600216.

Parker SJ, Jessel S, Richardson JE, et al. (2013) Older adults are mobile too! Identifying the barriers and facilitators to older adults' use of mHealth for pain management. *BMC geriatrics* 13: 43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3673892&tool=pmcentrez&rendertype=abstract.

Patomäki H and Wight C (2000) After postpositivism? The promises of critical realism. *International Studies Quarterly* 44(2): 213–237.

Pealer LN and Dorman SM (1997) Evaluating health-related web sites. 67(6): 232–235.

Pears S, Bijker M, Morton K, et al. (2016) A randomised controlled trial of three very brief interventions for physical activity in primary care. *BMC Public Health*, BMC Public Health 16(1): 1–13. Available from: http://dx.doi.org/10.1186/s12889-016-3684-7.

Peiris D, Praveen D and Johnson C (2014) Use of mHealth systems and tools for non-communicable diseases in low- and middle-income countries: a systematic review. *Journal of Cardiovascular Translational Research* 7(8): 677–91.

Pender NJ (2011) Heath promotion model manual.: 1–17. Available from: http://deepblue.lib.umich.edu/handle/2027.42/85350.

Peng W, Kanthawala S, Yuan S, et al. (2016) A qualitative study of user perceptions of mobile health apps. *BMC Public Health*: 1–11. Available from: http://dx.doi.org/10.1186/s12889-016-3808-0.

Pentakota N, Ramaswamy G, Thekkur P, et al. (2019) Is a smartphone application effective in improving physical activity among medical school students? Results from a quasi-experimental study. *International Journal of Adolescent Medicine and Health*.

Pentland A, Lazer D and Brewer D (2013) Improving Public Health and medicine by use of reality mining. *Social Science & Medicine* 3(7): 1–29.

Perski O, Blandford A, West R, et al. (2017) Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational Behavioral Medicine* 7(2): 254–267.

Peters D, Deady M, Glozier N, et al. (2018) Worker preferences for a mental health app within male-dominated industries: Participatory study. *Journal* *of Medical Internet Research* 20(4): 1–11.

Pew Research Centre (2017) *Mobile Fact Sheet*. *Pew Research Center: Internet, Science & Tech*. Available from: http://www.pewinternet.org/fact-sheet/mobile/.

Pipitprapat W, Harnchoowong S, Suchonwanit P, et al. (2018) The validation of smartphone applications for heart rate measurement. *Annals of Medicine* 50(8): 721–727.

Pisuwala U (n.d.) How often should you update your mobile app? *Peerbits*. Available from: https://www.peerbits.com/blog/update-mobile-app.html (accessed 2 August 2019).

Podina IR, Fodor LA, Cosmoiu A, et al. (2017) An evidence-based gamified mHealth intervention for overweight young adults with maladaptive eating habits: Study protocol for a randomized controlled trial. *Trials* 18(1): 1–14.

Podina IR, Faur AL, Fodor LA, et al. (2018) Usability and user experience testing of the cognitive-behavioral Sigma smartphone app for weight management. *Journal of Evidence-Based Psychotherapies* 18(1): 45–55.

Poojary J, Arora E, Britto A, et al. (2018) Validity of mobile-based technology vs direct observation in measuring number of steps and distance walked in 6 minutes. *Mayo Clinic Proceedings*, Mayo Foundation for Medical Education and Research 93(12): 1873–1874. Available from: https://doi.org/10.1016/j.mayocp.2018.09.003.

Popay J et al. (2006) Guidance on the conduct of narrative synthesis in systematic reviews. In: Results of an ESRC funded research project. (February 2016): 1–92.

Poushter J (2016) Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center*: 1–5. Available from: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/.

Powell AC, Torous J, Chan S, et al. (2016) Interrater reliability of mHealth app rating measures: analysis of top depression and smoking cessation apps. *JMIR mHealth and uHealth* 4(1): e15. Available from: http://mhealth.jmir.org/2016/1/e15/.

Pratt M, Sarmiento OL, Montes F, et al. (2012) The implications of megatrends in information and communication technology and transportation for changes in global physical activity. *Lancet (London, England)* 380(9838): 282–293. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=22818940&site=ehost-live.

Prestwich A, Webb TL and Conner M (2015) Using theory to develop and test interventions to promote changes in health behaviour: Evidence, issues, and recommendations. *Current Opinion in Psychology*, Elsevier Ltd 5(6): 1–5. Available from: http://dx.doi.org/10.1016/j.copsyc.2015.02.011.

Public Health England (2018) *Improving people's health: Applying behavioural and social sciences to improve population health and wellbeing in England*.

Pucci GCMF, Rech CR, Fermino RC, et al. (2012) Association between physical activity and quality of life in adults. *Revista de saúde pública* 46(1): 166–79.

Quantified Self Institute (n.d.) What is quantified self? Available from: http://qsinstitute.com/about/what-is-quantified-self/ (accessed 29 July 2019).

Rabin C and Bock B (2011) Desired features of smartphone applications promoting physical activity. *Telemedicine Journal and e-health* 17(10): 801–3. Available from: http://online.liebertpub.com/doi/abs/10.1089/tmj.2011.0055 (accessed 15 November 2014).

Raines C (2013) In-app mobile advertising : Investigating consumer attitudes towards pull- based mobile advertising amongst young adults in the UK. *Journal of Promotional Communications* 1 (1): 125–148.

Ramirez ER, Patrick K, Bietz M, et al. (2016) Discretionary self-monitoring of physical activity: A mixed-methods study of behavior change technique use and historical physical activity. *Public Health (Health Behavior)*. Available from: https://media.proquest.com/media/pq/classic/doc/4095709081/fmt/ai/rep/NPDF?_s=vj4F2yf09Jxn8nOSfh%2BkTBJNfWI%3D.

Ratner C (2012a) Objectivism. In: Given LM (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, Thousand Oaks: SAGE Publications, pp. 431–433.

Ratner C (2012b) Subjectivism. In: Given LM (ed.), *The SAGE Encyclopedia of Qualitative Research Methods By:*, Thousand Oaks: SAGE Publications, pp. 431–433.

Reavley NJ and Jorm AF (2012) Public recognition of mental disorders and beliefs about treatment: changes in Australia over 16 years. *The British Journal of Psychiatry* 200(5): 419–425. Available from: http://bjp.rcpsych.org/cgi/doi/10.1192/bjp.bp.111.104208.

Redfern J, Santo K, Coorey G, et al. (2016) Factors influencing engagement, perceived usefulness and behavioral mechanisms associated with a text message support program. *PLoS ONE* 11(10): 1–17. Available from: http://dx.doi.org/10.1371/journal.pone.0163929.

Renner B, Villinger K, Wahl DR, et al. (2018) Happy eater: A mobile intervention for boosting experienced eating. In: *32nd Conference of the European Health Psychology Society: Health Psychology Across the Lifespan: Uniting Research, Practice and Policy. 21-25 August 2018*, Galway, Ireland.

Renner B, Schupp H, King L, et al. (n.d.) SMART FOOD. *SMARTACT*. Available from: https://www.uni-konstanz.de/smartact/teilprojekte/ (accessed 26 July 2019).

Reynoldson C, Stones C, Allsop M, et al. (2014) Assessing the quality and usability of smartphone apps for pain self-management. *Pain Medicine (United States)* 15(6): 898–909.

Rhodes RE and Kates A (2015) Can the affective response to exercise predict future motives and physical activity behavior? A systematic review of published evidence. *Annals of Behavioral Medicine* 49(5): 715–731. Available from: http://link.springer.com/10.1007/s12160-015-9704-5.

Rhodes RE and Quinlan A (2015) Predictors of physical activity change among adults using observational designs.: 423–441.

Rhodes RE, Gray SM and Husband C (2019) Experimental manipulation of affective judgments about physical activity: A systematic review and meta-analysis of adults. *Health Psychology Review* 13(1): 18–34. Available from: https://www.tandfonline.com/doi/full/10.1080/17437199.2018.1530067.

Rhodes W (2019) 28 metrics that matter for your app. *Savvy Apps*. Available from: https://savvyapps.com/blog/mobile-app-analytics (accessed 29 July 2019).

Roberts AL, Fisher A, Smith L, et al. (2017) Digital health behaviour change interventions targeting physical activity and diet in cancer survivors: a systematic review and meta-analysis. *Journal of Cancer Survivorship* 11(6): 704–719.

Rocheleau CA, Webster GD, Bryan A, et al. (2004) Moderators of the relationship between exercise and mood changes: Gender, exertion level, and workout duration. *Psychology and Health* 19(4): 491–506.

Rose G (1985) Sick individuals and sick populations. *International Journal of Epidemiology* 14(1): 32–38.

Rose T, Barker M, Maria Jacob C, et al. (2017) A systematic review of digital interventions for improving the diet and physical activity behaviors of adolescents. *Journal of Adolescent Health* 61(6): 669–677.

Roth WR, Vilardaga R, Wolfe N, et al. (2014) Practical considerations in the design and development of smartphone apps for behavior change. *Journal of Contextual Behav Sci* 3(4): 269–272.

Rothman AJ (2000) Toward a theory-based analysis of behavioural maintenance. *Health Psychology* 19(1 (suppl)): 64–69.

Runyan J, Steenbergh T, Bainbridge C, et al. (2013) A smartphone ecological momentary assessment/intervention ‘app’ for collecting real-time data and promoting self awareness. *PLoS ONE* 8(8). Available from: http://www.plosone.org/article/fetchObject.action?uri=info%253Adoi%252F10.1371%252Fjournal.pone.0071325&representation=PDF (accessed 14 September 2014).

Russell J, Weiss A and Mendelsohn G (1989) Affect grid: a single-item scale of pleasure and arousal. *Journal Of Personality And Social Psychology* 57(3): 493–502.

Russell JA (2003) Core affect and the psychological construction of emotion. *Psychological Review* 110(1): 145–172.

Ryan RM and Deci EL (2001) On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annual Review Of Psychology* 52: 141–66.

Saadatfard O and Årsand E (2016) *M-health apps by numbers*. *Norwegian centre for health research. Fact sheets*.

Sala M, Baldwin AS and Williams DM (2016) Affective and cognitive predictors of affective response to exercise: Examining unique and overlapping variance. *Psychology of Sport and Exercise*, Elsevier Ltd 27: 1–8. Available from: http://dx.doi.org/10.1016/j.psychsport.2016.07.005.

Saleem MA, Fatima I, Khan KU, et al. (2012) Trajectory based activity monitoring and healthcare provisioning. In: *The Tenth IEEE International Conference on Pervasive Intelligence and Computing (PiCom 2012)*

Changzhou, China. Available from: http://uclab.khu.ac.kr/resources/publication/C_266.pdf%5Cnpapers3://publication/uuid/EC44DEDC-B915-41F1-A05F-7E090D1BB704.

Sallis JF and Saelens BE (2000) Assessment of physical activity by self-report: Status, limitations, and future directions. *Research Quarterly for Exercise and Sport* 71: 1–14.

Sallis JF, Johnson MF, Calfas KJ, et al. (1997) Assessing perceived physical environmental variables that may influence physical activity. *Research Quarterly for Exercise and Sport* 68(4): 345–351. Available from: http://www.tandfonline.com/doi/abs/10.1080/02701367.1997.10608015.

Salmon J, Owen N, Crawford D, et al. (2003) Physical activity and sedentary behavior: A population-based study of barriers, enjoyment, and preference. *Health Psychology* 22(2): 178–188.

Samdal GB, Eide GE, Barth T, et al. (2017) Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. *International Journal of Behavioral Nutrition and Physical Activity* 14(1): 1–14.

Sandelowski M, Voils C and Barroso J (2006) Defining and designing mixed research synthesis studies. *NIH Research in Schools (PMC)* 13(1): 29–40.

Sarcona A, Kovacs L, Wright J, et al. (2017) Differences in eating behavior, physical activity, and health-related lifestyle choices between users and nonusers of mobile health apps. *American Journal of Health Education*, Routledge 48(5): 298–305. Available from: https://doi.org/10.1080/19325037.2017.1335630.

Schensul JJ (2018) Methods. In: Given LM (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, Thousand Oaks: SAGE Publications, pp. 667–670.

Schneider JK, Eveker A, Bronder DR, et al. (2003) Exercise training program for older adults. Incentives and disincentives for participation. *Journal Of Gerontological Nursing* 29(9): 21–31. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=14528746&site=ehost-live.

Schneider ML and Kwan BM (2013) Psychological need satisfaction, intrinsic motivation and affective response to exercise in adolescents. *Psychology of Sport & Exercise* 14(5): 776–785.

Schoeppe S, Alley S, Rebar AL, et al. (2017) Apps to improve diet, physical activity and sedentary behaviour in children and adolescents: A review of quality, features and behaviour change techniques. *International Journal of Behavioral Nutrition and Physical Activity* 14(1): 1–10.

Schöndube A, Kanning M and Fuchs R (2016) The bidirectional effect between momentary affective states and exercise duration on a day level. *Frontiers in Psychology* 7(SEP).

Schuch FB, Vancampfort D, Rosenbaum S, et al. (2016) Exercise for depression in older adults: A meta-analysis of randomized controlled trials adjusting for publication bias. *Revista Brasileira de Psiquiatria* 38(3): 247–254.

Schutzer KA and Graves BS (2004) Barriers and motivations to exercise in older adults. *Preventive medicine* 39(5): 1056–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15475041 (accessed 9 July 2014).

Schwandt TA, Lincoln YS and Guba EG (2007) Judging interpretations: But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Evaluation* (114): 68–81.

Schwaneberg T, Weymar F, Ulbricht S, et al. (2017) Relationship between objectively measured intensity of physical activity and self-reported enjoyment of physical activity. *Preventive Medicine Reports* 7: 162–168. Available from: http://dx.doi.org/10.1016/j.pmedr.2017.06.004.

Schwerdtfeger A, Eberhardt R, Chmitorz A, et al. (2010) Momentary affect predicts bodily movement in daily life: an ambulatory monitoring study.

Journal of sport & exercise psychology 32(5): 674–693.

Scott K, Richards D and Adhikari R (2016) A review and comparative analysis of security risks and safety measures of mobile health apps. *Australasian Journal of Information Systems* 20(HealthCareBusinessTech 2014): 1–18.

Seefeldt V, Malina RM and Clark MA (2002) Factors affecting levels of physical activity in adults. *Sports Medicine* 32(3): 143–168.

Segar ML and Richardson CR (2014) Prescribing pleasure and meaning: Cultivating walking motivation and maintenance. *American Journal of Preventive Medicine*, Elsevier 47(6): 838–841. Available from: http://dx.doi.org/10.1016/j.amepre.2014.07.001.

Segar ML, Eccles JS and Richardson CR (2011) Rebranding exercise: closing the gap between values and behavior. *International Journal of Behavioral Nutrition and Phsyical Activity* 8(94): 1–14.

Shea BJ, Grimshaw JM, Wells GA, et al. (2007) Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC medical research methodology* 7: 10. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1810543&tool=pmcentrez&rendertype=abstract (accessed 4 March 2013).

Short CE, James EL, Plotnikoff RC, et al. (2011) Efficacy of tailored-print interventions to promote physical activity: a systematic review of randomised trials. *The International Journal of Behavioral Nutrition and Physical Activity*, BioMed Central Ltd 8(1): 113. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3214130&tool=pmcentrez&rendertype=abstract (accessed 4 June 2013).

Siegel PZ, Brackbill RM and Heath GW (1995) The epidemiology of walking for exercise: Implications for promoting activity among sedentary groups. *American Journal of Public Health* 85(5): 706–710.

Silberg WM (2011) Assessing, controlling, and assuring the quality of medical information on the internet. *JAMA* 277(15): 1244.

Sirriyeh R, Lawton R and Ward J (2010) Physical activity and adolescents : An exploratory randomized controlled trial investigating the influence of affective and instrumental text messages. *British Journal of Health Psychology* 15: 825–840.

Sirriyeh R, Lawton R, Gardner P, et al. (2012) Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice* 18(4): 746–752.

Skinner BF (1938) *The behavior of organisms: An experimental analysis.* New York: Appleton-Century.

Smith JK (2012) Relativism. In: Given LM (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, Thousand Oaks: SAGE Publications, pp. 667–670.

Speranzini N (2015) Do changes in affective responses during physical activity predict future physical activity behaviour in older adults? University of Ottawa.

Sport England (2015) This Girl Can. Available from: http://www.thisgirlcan.co.uk/.

Stadler G, Oettingen G and Gollwitzer PM (2009) Physical activity in women: effects of a self-regulation intervention. *American Journal of Preventive Medicine*, American Journal of Preventive Medicine 36(1): 29–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18977113 (accessed 14 November 2012).

Staten LK, Gregory-mercado KY, Ranger-moore J, et al. (2004) Provider counseling, health education, and community health workers: The Arizona WISEWOMAN Project. *Journal of Womens Health* 13(5): 547–556. Available from: http://www.liebertpub.com/media/content/jwh_wisewoman/p547.pdf.

Statista (2017) Global mobile OS market share in sales to end users from 1st quarter 2009 to 1st quarter 2017. *Statista*: 1–3. Available from: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/.

Statista (2018) *Most popular Apple App Store categories in September 2018, by share of available apps*. Available from: https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/.

Statista (n.d.) *Number of smartphone users worldwide from 2014 to 2020 (in billions)*. Available from: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

Steinhubl SR, Muse ED, Topol EJ, et al. (2016) The emerging field of mobile health. 7(283): 1–12.

Stephens J and Allen J (2013) Mobile phone interventions to increase physical activity and reduce weight: a systematic review. *The Journal of Cardiovascular Nursing* 28(4): 320–329.

Stevens CJ and Bryan AD (2012) Rebranding exercise: there's an app for that. *American Journal of Health Promotion* 27(2): 69–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23551199.

Stoyanov SR, Hides L, Kavanagh DJ, et al. (2015) Mobile App Rating Scale: A new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth* 3(1): e27. Available from: http://mhealth.jmir.org/2015/1/e27/.

Stoye G (2018) Does the NHS need more money and how could we pay for it? In: Kelly E, Johnson P, and Stoye G (eds), *The NHS at 70*, London. Available from: https://www.ifs.org.uk/uploads/HEAJ6320-Report-3-Does-The-NHS-Need-More-Money-180625-WEB.pdf.

Stralen V (2011) Determinants of initiation and maintenance of PA among older adults. *Health Psychology Review* 3(2).

Strath SJ, Kaminsky LA, Ainsworth BE, et al. (2013) Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the American Heart Association. *Circulation* 128(20): 2259–2279.

Stubbs B, Vancampfort D, Smith L, et al. (2018) Physical activity and mental health. *The Lancet Psychiatry*, Elsevier Ltd 5. Available from: http://dx.doi.org/10.1016/S2215-0366(18)30343-2.

Sylvia LG, Bernstein EE, Hubbard JL, et al. (2015) A practical guide to measuring physical activity. *Journal of the Academy of Nutrition and Dietetics* 114(2): 199–208.

Taki S, Denney-Wilson E, Laws R, et al. (2015) Infant feeding websites and apps: A systematic assessment of quality and content. *Interactive Journal of Medical Research* 4(3): e18.

Tanaka E, Osawa Y, Muramatsu K, et al. (2016) Evaluation of a device that promotes walking using a two-dimensional emotion map. In: *World Automation Congress*, pp. 1–5.

Taylor N (2014) Challenges in measuring physical activity in the context of mental health. In: Clow A and Edmunds S (eds), *Physical Activity and Mental Health*, Leeds: Human Kinetics, pp. 41–61.

Taylor S, Sano A, Ferguson C, et al. (2018) Quantifyme: An open-source automated single-case experimental design platform. *Sensors (Switzerland)* 18(4).

Teddlie C and Tashakkori A (2009) *Foundations of Mixed Methods Research*. London: Sage Publications Ltd.

The Theory and Techniques Tool (n.d.). Available from: https://theoryandtechniquetool.humanbehaviourchange.org/ (accessed 8 July 2019).

The WOOP App (n.d.). Available from: http://woopmylife.org/app.

Thompson Coon J, Boddy K, Stein K, et al. (2011) Does participating in physical activity in outdoor natural environments have a greater effect on physical and mental wellbeing than physical activity outdoors? A systematic review. *Environmental Science and Technology* 45: 1761–1772.

Thoren P, Floras JS, Hoffmann P, et al. (1990) Endorphins and exercise: physiological mechanisms and clinical implications. *Medicine and Science in Sports and Exercise* 22(4): 417–428.

Thorndike EL (1898) Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied* 2(4): 1–109.

Tollmar K, Bentley F and Viedma C (2012) Mobile Health Mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, IEEE: 65–72.

Tomita MR, Tsai B-M, Fisher NM, et al. (2008) Effects of multidisciplinary Internet-based program on management of heart failure. *Journal of Multidisciplinary Healthcare* 2009(2): 13–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2875686&tool=pmcentrez&rendertype=abstract.

Tong HL, Coiera E and Laranjo L (2018) Using a mobile social networking app to promote physical activity: a qualitative study of users' perspectives. *Journal of Medical Internet Research* 20(12): e11439.

Torous J, Nicholas J, Larsen ME, et al. (2018) Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. *Evidence-Based Mental Health* 21(3): 1–4.

Towler C, Fulk G, LaRue J, et al. (2018) Accuracy of 6 activity monitors during the 6 minute walk test to measure distance and steps walked. *Cardiopulmonary Physical Therapy Journal* 29(1).

Trost SG, Owen N, Bauman AE, et al. (2002) Correlates of adults participation in physical activity: Review and update. *Medicine & Science in Sports & Exercise*: 1996–2001.

Turner-McGrievy GM, Hales SB, Schoffman DE, et al. (2017) Choosing between responsive-design websites versus mobile apps for your mobile behavioral intervention: presenting four case studies. *Translational Behavioral Medicine* 7(2): 224–232.

Tuson KM and Sinyor D (1993) On the affective benefits of acute aerobic exercise. In: Seraganian P (ed.), *Exercise psychology: The influence of physical exercise on pyschological processes*, John Wiley & Sons, pp. 80–121.

Ullrich J, Cropper J, Frühwirt P, et al. (2016) The role and security of firewalls in cyber-physical cloud computing. *Eurasip Journal on Information Security* 2016(1). Available from: http://dx.doi.org/10.1186/s13635-016-0042-3.

Van Cappellen P, Rice EL, Catalino LI, et al. (2017) Positive affective processes underlie positive health behaviour change. *Psychology & Health* (May): 1–21. Available from: https://www.tandfonline.com/doi/full/10.1080/08870446.2017.1320798.

Van Gemert-Pijnen JEWC, Nijland N, Van Limburg M, et al. (2011) A holistic framework to improve the uptake and impact of eHealth technologies. *Journal of Medical Internet Research* 13(4).

Van Velthoven MH, Smith J, Wells G, et al. (2018) Digital health app development standards: A systematic review protocol. *BMJ Open* 8(8): 1–5.

Vandelanotte C and De Bourdeaudhuij I (2003) Acceptability and feasibility of a computer-tailored physical activity intervention using stages of change: project FAITH. *Health Education Research*, Oxford University Press 18(3): 304–317. Available from: https://academic.oup.com/her/article-lookup/doi/10.1093/her/cyf027 (accessed 28 November 2017).

Varpio L, Ajjawi R, Monrouxe L V., et al. (2017) Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. *Medical Education* 51(1): 40–50.

Vasa R, Hoon L, Mouzakis K, et al. (2012) A preliminary analysis of mobile app user reviews. *Proceedings of the 24th Australian Computer-Human Interaction Conference. ACM*: 241–244. Available from: http://dl.acm.org/citation.cfm?id=2414577.

Voicu RA, Dobre C, Bajenaru L, et al. (2019) Human physical activity recognition using smartphone sensors. *Sensors (Switzerland)* 19(3): 1–18.

Wahl DR, Villinger K, König LM, et al. (2017) Healthy food choices are happy food choices: Evidence from a real life sample using smartphone based assessments. *Scientific Reports* 7(1): 1–8.

Wakefield MA, Loken B and Hornik RC (2010) Use of mass media campaigns to change health behaviour. *The Lancet* 376(9748): 1261–1271.

Wallace BC, Small K, Brodley CE, et al. (2012) Deploying an interactive machine learning system in an evidence-based practice center. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*: 819. Available from: http://dl.acm.org/citation.cfm?doid=2110363.2110464.

Wardle J and Steptoe A (2003) Socioeconomic differences in attitudes and beliefs about healthy lifestyles. *J Epidemiol Community Health* 57(6): 440–443. Available from: http://www.hubmed.org/display.cgi?uids=12775791.

Webb TL, Joseph J, Yardley L, et al. (2010) Using the Internet to promote health behavior change: A systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of Medical Internet Research* 12(1).

West JH, Hall PC, Hanson CL, et al. (2012) There's an app for that: Content analysis of paid health and fitness apps. *Journal of Medical Internet*

Research 14(3): 1–12.

West JH, Hall PC, Arredondo V, et al. (2013) Health behavior theories in diet apps. *Journal of Consumer Health on the Internet* 17(1): 10–24.

White M, Adams J and Heywood P (2018) How and why do interventions that increase health overall widen inequalities within populations? *Social inequality and public health*: 65–82.

Wiles N, Tallon D, Stawarz K, et al. (2018) User experience of Cognitive Behavioral Therapy apps for depression: An analysis of app functionality and user reviews. *Journal of Medical Internet Research* 20(6): e10120.

Williams DM (2008) Exercise, affect, and adherence: An integrated model and a case for self-paced exercise. *Journal of sport & exercise psychology* 30(5): 471–496.

Williams DM, Dunsiger S, Ciccolo JT, et al. (2008) Acute affective response to a moderate-intensity exercise stimulus predicts physical activity participation 6 and 12 months later. *Psychology of Sport and Exercise* 9(3): 231–245.

Williams DM, Dunsiger S, Jennings EG, et al. (2012) Does affective valence during and immediately following a 10-min walk predict concurrent and future physical activity? *Annals of Behavioral Medicine* 44(1): 43–51.

Williams ED, Stamatakis E, Chandola T, et al. (2011) Assessment of physical activity levels in South Asians in the UK: findings from the Health Survey for England. *Journal of Epidemiology & Community Health* 65(6): 517–521. Available from: http://jech.bmj.com/cgi/doi/10.1136/jech.2009.102509.

Wolf G (n.d.) Antephase. Available from: http://antephase.com/bio (accessed 29 July 2019).

World Health Organisation (2010) *Global recommendations on physical activity for health*. Available from: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf%5Cn http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Global+Recomendations+on+physical+activity+for+health#0.

World Health Organisation (2011) mHealth: New horizons for health through mobile technologies. *Observatory* 3(June): 66–71. Available from: http://www.webcitation.org/63mBxLED9.

World Health Organisation (2012) Management of patient information. *Global Observatory for eHealth series* 6: 80.

World Health Organisation (2018) *Global action plan on physical activity 2018-2030: more active people for a healthier world*. Geneva.

World Health Organisation (n.d.) eHealth. Available from: https://www.who.int/ehealth/en/.

Yardley L, Spring BJ, Riper H, et al. (2016) Understanding and promoting effective engagement with digital behavior change interventions. *American Journal of Preventive Medicine*, Elsevier 51(5): 833–842. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0749379716302434.

Yarmosh K (2016) How often should you update your app? *Savvy Apps*. Available from: https://savvyapps.com/blog/how-often-should-you-update-your-app (accessed 2 August 2019).

Yasini M, Beranger J, Desmarais P, et al. (2016) mHealth quality: A process to seal the qualified mobile health apps. *Exploring Complexity in Health: An Interdisciplinary Systems Approach* 288: 205–209. Available from: http://ebooks.iospress.com/volumearticle/44601.

Yik M, Russell JA and Steiger JH (2011) A 12-point circumplex structure of core affect. *Emotion* 11(4): 705–731.

Zenko Z, Ekkekakis P and Ariely D (2016) Can you have your vigorous exercise and enjoy it too? Ramping intensity down increases postexercise, remembered, and forecasted pleasure. *Journal of Sport and Exercise Psychology* 38(2): 149–159. Available from:

http://journals.humankinetics.com/doi/10.1123/jsep.2015-0286.

Zhao J, Freeman B and Li M (2016) Can mobile phone apps influence people’s health behavior change? An evidence review. *Journal of Medical Internet Research* 18(11).

Zhou Y, Kankanhalli A and Huang K-W (2016) Effects of fitness applications with SNS: How do they influence physical activity research-in-progress. *Proceedings of the 37th International Conference on Information Systems* (2): 1–11. Available from: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1324&context=icis2016 .

Zolotareva K (2017) How to do ASO for Android Apps on Google Play Store - Ultimate Guide for 2019. *The Tool*. Available from: https://thetool.io/2017/how-to-optimize-google-play-aso-guide (accessed 29 July 2019).

Appendices

Appendix 1 Theory coding scheme framework and guidelines

Verbatim from (Michie and Prestwich, 2010)

Theory Coding Scheme 12e: Instructions

The coding scheme that follows comprises 19 items. For each item, code as Yes or No. When 'Yes', state the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).

The coding scheme is based on the *explicit* use of theory. Consequently, even when theory-relevant information is implied but is not stated explicitly, the related items should be coded as 'no'.

Defining Terms

Please refer to the definitions (see below) during coding:

Theory (or Model)

'a set of interrelated concepts, definitions and propositions that present a *systematic* view of events or situations by specifying relations among variables, in order to *explain* or *predict* the events or situations' (Glanz & Rimer, 1995). Examples include: TPB, TRA, HBM, stage of change/trans-theoretical model etc.)

Theory-relevant construct

A construct (a key concept, excluding behaviour) within a theory/model upon which the intervention is based. Please refer to the '*Table of Theories*' to identify whether a particular construct belongs to the specified theory.

Predictors

Constructs that are not explicitly linked to a theory by the authors, but are targeted for intervention (as a means to change behaviour) because they predict behaviour. Predictors must only be coded if the author has presented evidence that the construct predicts/correlates with/causes behaviour. Predictors do not include actual behaviour, self-reported or otherwise (e.g. amount of time spent exercising), or biological factors (e.g., age, sex).

Intervention Technique

Strategy used to change behaviour, theory-relevant construct or predictor (e.g., providing information on consequences; prompting specific goal setting; prompting barrier identification; modelling the behaviour; planning social support).

Item Specific Information

i. Items 4-13 and 15-16: Code theory-relevant constructs and predictors separately.

ii. Items 17-19: Only code according to the theory base of the intervention (i.e. not predictors)

iii. Items 7-9: In some cases, items 7, 8 and 9 may all be coded as 'no' (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa).

If item 7 is coded 'yes' then items 8 and 9 must be coded 'no'.

It is possible for items 8 and 9 to both be coded as 'yes'.

In instances where the name of the *construct* might overlap with the name of the technique (e.g., tips on *social support* were given; exercise *benefits* were discussed), code this as a direct link between technique and construct (i.e. they do not need to be written in the form 'tips on social support [technique] were used to target social support [construct]'; exercise benefits were discussed [technique] to target perceived benefits [construct]).

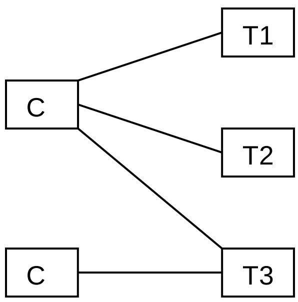
iv. Footnotes:

[1] *Item 7:* All intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C). Examples of meeting this criterion:

- T1 & T2 are used to change C1; T3 is used to change C1 and C2. Techniques 1-3 are the only intervention techniques used (see DIAGRAM 1).

- T1 used to change C1; T2 used to change C2; T3 used to change C3

DIAGRAM 1: Example: All intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)

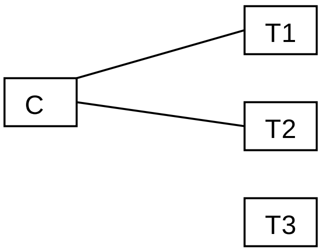


[2] *Item 8:* At least one, but not all, of the intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C). At least one technique, therefore, is not linked to any theory-relevant construct/predictor. Examples of meeting this criterion:

- T1 and T2 are used to change C1; T3 is used but not linked to a construct/predictor (see DIAGRAM 2).

- T1 used to change C1; T2 used to change C1 and C2; T3 is used but not linked to a construct/predictor.

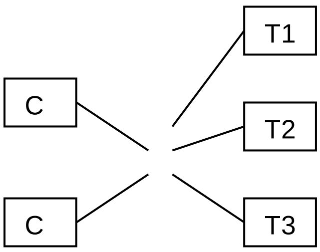
DIAGRAM 2: Example: At least one, but not all, of the intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)



[3] *Item 9:* Group of techniques (T) are linked to a group of theory-relevant constructs/predictors (C). Example of meeting this criterion:

- T1, T2, T3 are used to change C1, C2 (see DIAGRAM 3).

DIAGRAM 3: Example: Group of techniques (T) are linked to a group of theory-relevant constructs/predictors (C)



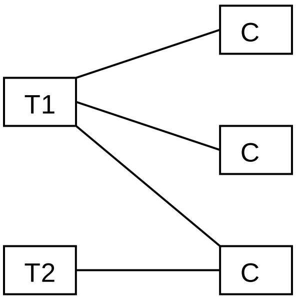
[4] *Item 10:* All constructs within a stated theory/all stated predictors (C) are linked to at least one intervention technique (T). Examples of meeting this criterion:

- C1 & C2 are linked to T1; C3 is linked to T2. Constructs 1-3 are the only constructs within the theory specified in item 5 (see DIAGRAM 4).

- C1 is linked to T1; C2 is linked to T2; C3 is linked to T3

To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the '*Table of Theories*'.

DIAGRAM 4: Example: All constructs within a stated theory/all predictors (C) are linked to at least one intervention technique (T)

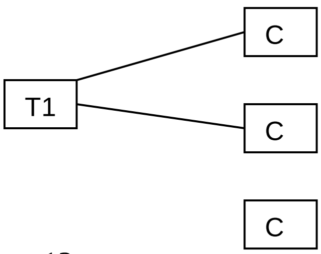


[5] *Item 11:* At least one, but not all, of the constructs within a stated theory/all stated predictors (C) are linked to at least one intervention technique (T). At least one construct within a stated theory, therefore, is not linked to any intervention technique. Examples of meeting this criterion:

> - C1 and C2 are linked to T1; C3 is not linked to an intervention technique- or is not highlighted by the authors (see DIAGRAM 5).
>
> - C1 is linked to T1; C2 is linked to T1 and T2; C3 is not linked to an intervention technique- or is not highlighted by the authors.

To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the '*Table of Theories*'.

DIAGRAM 5: Example: At least one, but not all, of the constructs within a stated theory/stated predictors (C) are linked to at least one intervention technique (T)



[6] *Item 13*

Reliability/Validity.

- To judge measures as valid authors should present reference to relevant analyses the authors have conducted *or* with reference to such analyses conducted by others (including via citation).

- To judge a measure as reliable, the authors should test whether their measure, for their sample, is reliable by conducting and reporting relevant analyses.

- For the purposes of developing the scheme, we have used the following criteria (alpha >.7; test-retest reliability >.7; reference to satisfactory validity).

- Single item measures are seen as reliable/valid only with appropriate supporting evidence (e.g., test-retest reliability >.7). However, given the various issues connected with single item measures, those using the scheme may wish to highlight studies that have incorporated single-item, but otherwise reliable/valid, measures (e.g., using *) to differentiate these from studies employing reliable/valid multi-item scales.

- Where authors have used multiple measures (or sub-scales) and these have subsequently been combined into a single index (e.g., for the purpose of analysis), the authors should cite evidence for the reliability/validity of the overall index and/or for each of the component measures.

- Where authors have used multiple measures and these have not been combined into a single index (e.g., for the purpose of analysis) then the authors should cite evidence for the reliability/validity of at least one of the measures.

- Those using the Theory Coding Scheme for review purposes may wish to alternatively define the parameters required to demonstrate satisfactory reliability and validity,

[7] Items 13a), 13b), 13c), 13d)

Where appropriate evidence for at least one type of reliability/validity is reported, items 13a) (reliability) and 13c) (validity) should be coded ‘yes’ and items 13b) (reliability) and 13d) (validity) should be coded ‘no’.

[8] *Item 14*

Randomization.

- In instances where authors report randomization checks that demonstrate equality across experimental conditions (or inequalities between experimental conditions are statistically controlled) but they have excluded at least one of the measures that could have been included in randomization checks, this should still be coded as an instance of successful randomization.

- Where groups are allocated to condition, the authors should note how the groups were randomized to condition (e.g., coin toss).

- Method of randomization refers specifically to the method used to generate the random allocation sequence (e.g., using a random number generator) rather than types of randomization (e.g., stratified random sampling).

<u>Additional Notes:</u>

1. When more than one intervention is used (excluding the control group), the items should be coded separately for each intervention
2. Note for Stage of Change/Trans-Theoretical Model: While the construct 'stage of change' may either be used to select recipients of an intervention (item 4) or to tailor an intervention (item 6), 'stage of change' measures should not be coded for items 7-16. However, self-efficacy and decisional balance, both constructs within the TTM, should be coded throughout.

Paper ID: Coder: Date:

Item No.	Item	Description	Yes/ No/ Don't know	List with location	
1	*Theory/model of behaviour mentioned*	Models/theories that specify relations among variables, in order to *explain* or *predict* behaviour (e.g., TPB, SCT, HBM) are mentioned, even if the intervention is not based on this theory			
2	*Targeted construct mentioned as predictor of behaviour*	('Targeted' construct refers to a psychological construct that the study intervention is hypothesised to change). Evidence that the psychological construct relates to (correlates/predicts/causes) behaviour should be presented within the introduction or method (rather than the Discussion).		**Location of evidence that construct relates to behaviour:** **Location that this predictor is targeted by the intervention:**	
3	*Intervention based on single theory*	The intervention is based on a single theory (rather than a combination of theories or theory + predictors)			
4	*Theory/ predictors used to select recipients for the intervention*	Participants were screened/selected based on achieving a particular score/level on a theory-relevant construct/predictor		**Construct (Theory)**	**Predictor**
5	*Theory/ predictors used to select/develop intervention techniques*	The intervention is explicitly based on a theory or predictor or combination of theories or predictors		**Theory**	**Predictor**
6	*Theory/ predictors used to tailor intervention techniques to recipients*	The intervention differs for different sub-groups that vary on a psychological construct (e.g., stage of change) or predictor at baseline		**Construct**	**Predictor**
7	*All intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor*	Each intervention technique is explicitly linked to at least one theory-relevant construct/predictor.[1]		**Construct (list links)**	**Predictor (list links)**

8	*At least one, but not all, of the intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor*	At least one, but not all, of the intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor.[2]		**Construct (list links)**	**Predictor (list links)**
9	*Group of techniques are linked to a group of constructs/ predictors*	A cluster of techniques is linked to a cluster of constructs/ predictors.[3]		**List clusters of techniques/constructs**	**List clusters of techniques/predictors**
10	*All theory-relevant constructs/predictors are explicitly linked to at least one intervention technique*	Every theoretical construct within a stated theory, or every stated predictor (see item 5), is linked to at least one intervention technique.[4]		**Construct (list links)**	**Predictor (list links)**
11	*At least one, but not all, of the theory relevant constructs/predictors are explicitly linked to at least one intervention technique*	At least one, but not all, of the theoretical constructs within a stated theory or at least one, but not all, of the stated predictors (see item 5) are linked to at least one intervention technique.[5]		**Construct (list links)**	**Predictor (list links)**

12	*Theory-relevant constructs/ predictors are measured*	a) At least one construct of theory (or predictor) mentioned in relation to the intervention is measured POST-INTERVENTION. b) At least one construct of theory (or predictor) mentioned in relation to the intervention is measured PRE AND POST-INTERVENTION.		**Construct**	**Predictor**
13	*Quality of Measures*[6]	a) All of the measures of theory relevant constructs/predictors had some evidence for their reliability[7] b) At least one, but not all, of the measures of theory relevant constructs/predictors had some evidence for their reliability c) All of the measures of theory relevant constructs/predictors have been previously validated d) At least one, but not all, of the measures of theory relevant constructs/predictors have been previously validated e) The behaviour measure had some evidence for its reliability f) The behaviour measure has been previously validated		**Construct**	**Predictor**
14	*Randomization of participants to condition*[8]	a) Do the authors claim randomization? b) Is a method of random allocation to condition described (e.g., random number generator; coin toss) c) Was the success of randomization tested? d) Was the randomization successful (or baseline differences between intervention and control group statistically controlled)?			
15	*Changes in measured theory-*	The intervention leads to sig. change in at least one		**Construct**	**Predictor**

	relevant constructs/ predictor	theory-relevant construct/predictor (vs. control group) in favour of the intervention.			
16	*Mediational analysis of construct/s / predictors*	In addition to 15, do the following effects emerge?: a) Mediator predicts DV? (or change in mediator leads to change in DV) b) Mediator predicts DV (when controlling for IV)? c) Intervention does not predict DV (when controlling for mediator)? d) Mediated effect statistically significant?		**Construct**	**Predictor**
17	*Results discussed in relation to theory*	Results are discussed in terms of the theoretical basis of the intervention			
18	*Appropriate support for theory*	Support for the theory is based on appropriate mediation OR refutation of the theory is based on obtaining appropriate null effects (i.e. changing behaviour without changing the theory-relevant constructs).			
19	*Results used to refine theory*	The authors attempt to refine the theory upon which the intervention was based by either: a) adding or removing constructs to the theory, or b) specifying that the interrelationships between the theoretical constructs should be changed and spelling out which relationships should be changed		**a) Constructs added or removed from theory:** **b) Interrelationships between the theoretical constructs to be changed:**	

Table of Theories

Theory	***Theory-relevant constructs***
Theory of Reasoned Action	1. **Attitudes**; 2. **SN**; 3. **Intentions**
Theory of Planned Behaviour	1. **Attitudes**; 2. **SN**; 3. **PBC**; 4. **Intentions**
Trans-Theoretical Model/Stages of Change	1. **Self-Efficacy** (person's confidence in performing a particular behaviour); 2. **Decisional Balance**
SCT	1. **Self-Efficacy** (person's confidence in performing a particular behavior)/ **Behavioral capability** (Knowledge and skill to perform a given behavior); 2. **Action-Outcome Expectancies** (extent that one's actions are seen as instrumental for the outcome/values associated with outcomes)/**attitudes**; 3. **Barriers** (including changes to environment/emotional barriers or one's perceptions of them). The following constructs are also related to the theory (and subsequently should be listed when they are cited by the authors): *Behavioral capability:* Knowledge and skill to perform a given behavior; *Attitudes (outcome-expectancies):-* *Expectations:* Anticipatory outcomes of a behavior; *Expectancies:* The values that the person places on a given outcome, incentives; *Self-control:* Personal regulation of goal-directed behavior or performance *Goals?*
Health Belief Model	1. **Perceived Susceptibility**; 2. **Perceived Severity**; 3. **Perceived Benefits** (one's belief in the efficacy of the advised action to reduce risk or seriousness of impact); 4. **Perceived Barriers**; 5. **Cues to Action**; 6. **Self-Efficacy** (added by Rosenstock et al., 1988)
Protection Motivation Theory (PMT)	1. **Intention**; 2. **PBC**; 3. **Perceived Severity**; 4. **Perceived Vulnerability**; 5. **Response Efficacy**; 6. **Response Costs**; 7. **Fear** (now added to the model),
Rubicon Model / Model of Action Phases (Heckhausen, 1991; Gollwitzer, 1990)	1. **Motivation** (e.g., intention); 2. **Volition** (e.g., planning)

Mobile Application Rating Scale (MARS)

App Classification

The Classification section is used to collect descriptive and technical information about the app. Please review the app description in iTunes / Google Play to access this information.

App Name: ____________________

Rating this version: ____________ **Rating all versions:** ____________

Developer: ____________________

N ratings this version: ____________ **N ratings all versions:** ____________

Version: ____________ **Last update:** ____________

Cost - basic version: ____________ **Cost - upgrade version:** ____________

Platform: ☐ iPhone ☐ iPad ☐ Android

Brief description: ____________________

____

Focus: what the app targets (select all that apply)

- ☐ Increase Happiness/Well-being
- ☐ Mindfulness/Meditation/Relaxation
- ☐ Reduce negative emotions
- ☐ Depression
- ☐ Anxiety/Stress
- ☐ Anger
- ☐ Behaviour Change
- ☐ Alcohol /Substance Use
- ☐ Goal Setting
- ☐ Entertainment
- ☐ Relationships
- ☐ Physical health
- ☐ Other ____________

Theoretical background/Strategies (all that apply)

- ☐ Assessment
- ☐ Feedback
- ☐ Information/Education
- ☐ Monitoring/Tracking
- ☐ Goal setting
- ☐ Advice /Tips /Strategies /Skills training
- ☐ CBT - Behavioural (positive events)
- ☐ CBT – Cognitive (thought challenging)
- ☐ ACT - Acceptance commitment therapy
- ☐ Mindfulness/Meditation
- ☐ Relaxation
- ☐ Gratitude
- ☐ Strengths based
- ☐ Other ____________

Affiliations:

☐ Unknown ☐ Commercial ☐ Government ☐ NGO ☐ University

Age group (all that apply)

- ☐ Children (under 12)
- ☐ Adolescents (13-17)
- ☐ Young Adults (18-25)
- ☐ Adults
- ☐ General

Technical aspects of app (all that apply)

- ☐ Allows sharing (Facebook, Twitter, etc.)
- ☐ Has an app community
- ☐ Allows password-protection
- ☐ Requires login
- ☐ Sends reminders
- ☐ Needs web access to function

QUT Queensland University of Technology

YOUNGANDWELL Cooperative Research Centre

App Quality Ratings

The Rating scale assesses app quality on four dimensions. All items are rated on a 5-point scale from "1.Inadequate" to "5.Excellent". Circle the number that most accurately represents the quality of the app component you are rating. Please use the descriptors provided for each response category.

SECTION A

Engagement – fun, interesting, customisable, interactive (e.g. sends alerts, messages, reminders, feedback, enables sharing), well-targeted to audience

1. **Entertainment: Is the app fun/entertaining to use? Does it use any strategies to increase engagement through entertainment (e.g. through gamification)?**

   1 Dull, not fun or entertaining at all
   2 Mostly boring
   3 OK, fun enough to entertain user for a brief time (< 5 minutes)
   4 Moderately fun and entertaining, would entertain user for some time (5-10 minutes total)
   5 Highly entertaining and fun, would stimulate repeat use

2. **Interest: Is the app interesting to use? Does it use any strategies to increase engagement by presenting its content in an interesting way?**

   1 Not interesting at all
   2 Mostly uninteresting
   3 OK, neither interesting nor uninteresting; would engage user for a brief time (< 5 minutes)
   4 Moderately interesting; would engage user for some time (5-10 minutes total)
   5 Very interesting, would engage user in repeat use

3. **Customisation: Does it provide/retain all necessary settings/preferences for apps features (e.g. sound, content, notifications, etc.)?**

   1 Does not allow any customisation or requires setting to be input every time
   2 Allows insufficient customisation limiting functions
   3 Allows basic customisation to function adequately
   4 Allows numerous options for customisation
   5 Allows complete tailoring to the individual's characteristics/preferences, retains all settings

4. **Interactivity: Does it allow user input, provide feedback, contain prompts (reminders, sharing options, notifications, etc.)? Note: these functions need to be customisable and not overwhelming in order to be perfect.**

   1 No interactive features and/or no response to user interaction
   2 Insufficient interactivity, or feedback, or user input options, limiting functions
   3 Basic interactive features to function adequately
   4 Offers a variety of interactive features/feedback/user input options
   5 Very high level of responsiveness through interactive features/feedback/user input options

5. **Target group: Is the app content (visual information, language, design) appropriate for your target audience?**

   1 Completely inappropriate/unclear/confusing
   2 Mostly inappropriate/unclear/confusing
   3 Acceptable but not targeted. May be inappropriate/unclear/confusing
   4 Well-targeted, with negligible issues
   5 Perfectly targeted, no issues found

A. Engagement mean score = ______

Queensland University of Technology

YOUNGANDWELL
Cooperative Research Centre

SECTION B

Functionality – app functioning, easy to learn, navigation, flow logic, and gestural design of app

6. **Performance: How accurately/fast do the app features (functions) and components (buttons/menus) work?**

    1 App is broken; no/insufficient/inaccurate response (e.g. crashes/bugs/broken features, etc.)
    2 Some functions work, but lagging or contains major technical problems
    3 App works overall. Some technical problems need fixing/Slow at times
    4 Mostly functional with minor/negligible problems
    5 Perfect/timely response; no technical bugs found/contains a 'loading time left' indicator

7. **Ease of use: How easy is it to learn how to use the app; how clear are the menu labels/icons and instructions?**

    1 No/limited instructions; menu labels/icons are confusing; complicated
    2 Useable after a lot of time/effort
    3 Useable after some time/effort
    4 Easy to learn how to use the app (or has clear instructions)
    5 Able to use app immediately; intuitive; simple

8. **Navigation: Is moving between screens logical/accurate/appropriate/ uninterrupted; are all necessary screen links present?**

    1 Different sections within the app seem logically disconnected and random/confusing/navigation is difficult
    2 Usable after a lot of time/effort
    3 Usable after some time/effort
    4 Easy to use or missing a negligible link
    5 Perfectly logical, easy, clear and intuitive screen flow throughout, or offers shortcuts

9. **Gestural design: Are interactions (taps/swipes/pinches/scrolls) consistent and intuitive across all components/screens?**

    1 Completely inconsistent/confusing
    2 Often inconsistent/confusing
    3 OK with some inconsistencies/confusing elements
    4 Mostly consistent/intuitive with negligible problems
    5 Perfectly consistent and intuitive

B. Functionality mean score =

SECTION C

Aesthetics – graphic design, overall visual appeal, colour scheme, and stylistic consistency

10. **Layout: Is arrangement and size of buttons/icons/menus/content on the screen appropriate or zoomable if needed?**

    1 Very bad design, cluttered, some options impossible to select/locate/see/read. Device display not optimised
    2 Bad design, random, unclear, some options difficult to select/locate/see/read
    3 Satisfactory, few problems with selecting/locating/seeing/reading items or with minor screen-size problems
    4 Mostly clear, able to select/locate/see/read items
    5 Professional, simple, clear, orderly, logically organised, device display optimised. Every design component has a purpose

11. **Graphics: How high is the quality/resolution of graphics used for buttons/icons/menus/content?**

    1 Graphics appear amateur, very poor visual design - disproportionate, completely stylistically inconsistent
    2 Low quality/low resolution graphics; low quality visual design – disproportionate, stylistically inconsistent
    3 Moderate quality graphics and visual design (generally consistent in style)
    4 High quality/resolution graphics and visual design – mostly proportionate, stylistically consistent
    5 Very high quality/resolution graphics and visual design - proportionate, stylistically consistent throughout

12. **Visual appeal: How good does the app look?**

    1 No visual appeal, unpleasant to look at, poorly designed, clashing/mismatched colours
    2 Little visual appeal – poorly designed, bad use of colour, visually boring
    3 Some visual appeal – average, neither pleasant, nor unpleasant
    4 High level of visual appeal – seamless graphics – consistent and professionally designed
    5 As above + very attractive, memorable, stands out; use of colour enhances app features/menus

C. Aesthetics mean score = _____________

SECTION D

Information – Contains high quality information (e.g. text, feedback, measures, references) from a credible source. Select N/A if the app component is irrelevant.

13. **Accuracy of app description (in app store): Does app contain what is described?**

    1 Misleading. App does not contain the described components/functions. Or has no description
    2 Inaccurate. App contains very few of the described components/functions
    3 OK. App contains some of the described components/functions
    4 Accurate. App contains most of the described components/functions
    5 Highly accurate description of the app components/functions

14. **Goals: Does app have specific, measurable and achievable goals (specified in app store description or within the app itself)?**

    N/A Description does not list goals, or app goals are irrelevant to research goal (e.g. using a game for educational purposes)
    1 App has no chance of achieving its stated goals
    2 Description lists some goals, but app has very little chance of achieving them
    3 OK. App has clear goals, which may be achievable.
    4 App has clearly specified goals, which are measurable and achievable
    5 App has specific and measurable goals, which are highly likely to be achieved

15. **Quality of information: Is app content correct, well written, and relevant to the goal/topic of the app?**

    N/A There is no information within the app
    1 Irrelevant/inappropriate/incoherent/incorrect
    2 Poor. Barely relevant/appropriate/coherent/may be incorrect
    3 Moderately relevant/appropriate/coherent/and appears correct
    4 Relevant/appropriate/coherent/correct
    5 Highly relevant, appropriate, coherent, and correct

QUT Queensland University of Technology

YOUNGANDWELL Cooperative Research Centre

16. **Quantity of information: Is the extent coverage within the scope of the app; and comprehensive but concise?**

    N/A There is no information within the app
    1 Minimal or overwhelming
    2 Insufficient or possibly overwhelming
    3 OK but not comprehensive or concise
    4 Offers a broad range of information, has some gaps or unnecessary detail; or has no links to more information and resources
    5 Comprehensive and concise; contains links to more information and resources

17. **Visual information: Is visual explanation of concepts – through charts/graphs/images/videos, etc. – clear, logical, correct?**

    N/A There is no visual information within the app (e.g. it only contains audio, or text)
    1 Completely unclear/confusing/wrong or necessary but missing
    2 Mostly unclear/confusing/wrong
    3 OK but often unclear/confusing/wrong
    4 Mostly clear/logical/correct with negligible issues
    5 Perfectly clear/logical/correct

18. **Credibility: Does the app come from a legitimate source (specified in app store description or within the app itself)?**

    1 Source identified but legitimacy/trustworthiness of source is questionable (e.g. commercial business with vested interest)
    2 Appears to come from a legitimate source, but it cannot be verified (e.g. has no webpage)
    3 Developed by small NGO/institution (hospital/centre, etc.) /specialised commercial business, funding body
    4 Developed by government, university or as above but larger in scale
    5 Developed using nationally competitive government or research funding (e.g. Australian Research Council, NHMRC)

19. **Evidence base: Has the app been trialled/tested; must be verified by evidence (in published scientific literature)?**

    N/A The app has not been trialled/tested
    1 The evidence suggests the app does not work
    2 App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has partially positive outcomes in studies that are not randomised controlled trials (RCTs), or there is little or no contradictory evidence.
    3 App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has positive outcomes in studies that are not RCTs, and there is no contradictory evidence.
    4 App has been trialled and outcome tested in 1-2 RCTs indicating positive results
    5 App has been trialled and outcome tested in $\geq$ 3 high quality RCTs indicating positive results

D. Information mean score = ____________ *

* Exclude questions rated as "N/A" from the mean score calculation.

Queensland University of Technology

YOUNGANDWELL Cooperative Research Centre

App subjective quality

SECTION E

20. **Would you recommend this app to people who might benefit from it?**

1	**Not at all**	I would not recommend this app to anyone
2		There are very few people I would recommend this app to
3	**Maybe**	There are several people whom I would recommend it to
4		There are many people I would recommend this app to
5	**Definitely**	I would recommend this app to everyone

21. **How many times do you think you would use this app in the next 12 months if it was relevant to you?**

1	**None**
2	1-2
3	3-10
4	10-50
5	>50

22. **Would you pay for this app?**

1	No
3	Maybe
5	Yes

23. **What is your overall star rating of the app?**

1	★	One of the worst apps I've used
2	★★	
3	★★★	Average
4	★★★★	
5	★★★★★	One of the best apps I've used

Scoring

App quality scores for

SECTION

A: Engagement Mean Score = ______________________

B: Functionality Mean Score = ______________________

C: Aesthetics Mean Score =

D: Information Mean Score = ______________________

App quality mean Score = ______________________

App subjective quality Score = ____________________

Queensland University of Technology

YOUNGANDWELL Cooperative Research Centre

App-specific

These added items can be adjusted and used to assess the perceived impact of the app on the user's knowledge, attitudes, intentions to change as well as the likelihood of actual change in the target health behaviour.

SECTION F

1. **Awareness: This app is likely to increase awareness of the importance of addressing [insert target health behaviour]**

Strongly disagree				Strongly Agree
1	2	3	4	5

2. **Knowledge: This app is likely to increase knowledge/understanding of [insert target health behaviour]**

Strongly disagree				Strongly Agree
1	2	3	4	5

3. **Attitudes: This app is likely to change attitudes toward improving [insert target health behaviour]**

Strongly disagree				Strongly Agree
1	2	3	4	5

4. **Intention to change: This app is likely to increase intentions/motivation to address [insert target health behaviour]**

Strongly disagree				Strongly Agree
1	2	3	4	5

5. **Help seeking: Use of this app is likely to encourage further help seeking for [insert target health behaviour] (if it's required)**

Strongly disagree				Strongly Agree
1	2	3	4	5

6. **Behaviour change: Use of this app is likely increase/decrease [insert target health behaviour]**

Strongly disagree				Strongly Agree
1	2	3	4	5

Appendix 3 Systematic review data extraction guidelines

Data extraction and quality assessment guidelines

Systematic Review

General guidelines and information

Research questions

The purpose of this data extraction form is to collect data from eligible literature to answer the following research questions:

1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps that include feedback on immediate affect?
3. What is the quality of these physical activity apps, that include feedback on immediate affect?

Data extraction form and documents

The data extraction form consists of five excel sheets that allow for data extraction as well as quality assessment. Tabs include: Data Extraction, QA_TCS, QA_PA BCTs, QA_MARS and QA_PA measure reliability. This document will be divided into different sections per tab.

You will have been provided with:

- the paper describing the app
- the app on an appropriate device if available
- a link to the app in the app store if available
- other associated papers that provide evidence for the app if available

Where these are available ALL should be used to perform the data extraction.

You should ‘test’ the app over a 48 hour period to ensure you are aware of all functionality.

These guidelines include a number of appendices. Some are at the end of this document and some are in separate documents which you will have been provided with, due to their size. Please see the guidelines for where each appendix is located.

If any item is not reported, write: unreported, unless otherwise stated.

A: Data extraction tab

1. **Admin details**

Item	**Explanation**
1.1 Coder initials	Enter your initials
1.2 Date of extraction	Enter the date you started the data extraction
1.3 Endnote Record Number	This is the number that Endnote associated with the paper – it will either already be entered or be written on a paper copy of the paper for you
1.4 Author	First authors last name and et al, if more than one
1.5 Title	Full title of paper
1.6 Year	Year of publication
1.7 Country study conducted in	Country where study took place, be as specific as possible (e.g. England, not just UK). If information not reported put: Unreported
1.8 App availability	Yes/No If available, you should have been provided with a copy of the app on the relevant device.
1.9 Extra evidence provided?	Yes/No If available, you should have been provided with additional papers providing evidence/details of the app – specify here.
1.10 Additional notes	If you have any other concerns/thoughts about this section, put them here.

2. **Methods**

Item	**Explanation**
2.1 Study type	Qualitative, Quantitative, Mixed methods, Multi methods. Mixed methods = where methods/findings are integrated at some point either during data collection or analysis or synthesis, typically to answer the SAME question. Multi methods = Where multiple methods of data collection are used to answer different questions and therefore no integration of the two is performed at any time.
2.2 Study design	Report any and all designs used. Possible Quantitative designs (not exhaustive). See appendix 1 at end of this document for characteristics. Also, see section 9.3 of the Cochrane

	handbook for definitions of some study designs - • RCT=Randomized controlled trial • Q-RCT=Quasi-randomized controlled trial • NRCT=Non-randomized controlled trial • CBA=Controlled before-and-after study • PCS=Prospective cohort study • RCS=Retrospective cohort study • HCT=Historically controlled trial • NCC=Nested case-control study • CC=Case-control study • XS=Cross-sectional study • BA=Before-and-after comparison • CR/CS=Case report/Case series • ClRCT=Cluster randomized controlled trial • ClQ-RCT=Cluster quasi-randomized controlled trial • ClNRT=Cluster non-randomized controlled trial • CITS=Controlled interrupted time series • CChBA=Controlled cohort before-and-after study • ITS=Interrupted time series • ChBA=Cohort before and after study • EcoXS=Ecological cross-sectional study Possible Qualitative designs (not exhaustive). • Biographical study • Phenomenology • Grounded theory • Ethnography • Case study • Narrative • Historical • Action Research • Co-design – where an app is being developed based on qualitative information • Process evaluation If qualitative, also describe data collection methods (e.g. interviews, focus groups, observation, document analysis etc.) For descriptions of qualitative studies see: https://researchrundowns.com/qual/qualitative-

	research-design/ https://measuringu.com/qual-methods/ Also see Nieswiadomy, (2012) – appendix 2 (separate document) for more details on qualitative designs
2.3 Study duration	From consent to follow-up if available, otherwise any information related to duration that is provided for any of the sections of the study e.g. duration over which focus groups took place.
2.4 Additional notes	If you have any other concerns/thoughts about this section, put them here.

3. Participants using app

Item	Explanation
3.1 Sample size	Number of participants in study and per group, including details on attrition for different groups/sections of the study as applicable or reported.
3.2 Setting – recruitment AND implementation/testing	Home, community, school, hospital, lab, etc. – where were participants RECRUITED from and where did the study TAKE PLACE if different. If testing/using app in everyday life, then list setting as ‘community’. Do this for all parts of the study if there are multiple e.g. focus group settings AND pilot testing setting.
3.3 Diagnostic criteria	If any HEALTH/PSYCHOLOGICAL/PHYSICAL/ WELLBEING/PERSONALITY/ MOTIVATIONAL /BEHAVIOURAL criteria were used to screen participants, give details e.g. sedentary as defined by a specific PA self-report measure and units. Includes things such as smartphone ownership/technology familiarity.
3.4 Age	Participant age range and standard deviation if reported, AND if participation was limited based on age. Report for both recruitment criteria AND final sample if available and specify which is which. Provide whatever information is available in paper.
3.5 Sex	No. of males/females and if participation was limited based on sex

	Report for both recruitment criteria AND final sample if available and specify which is which. Provide whatever information is available in paper.
3.6 Country (if different from conducted)	Country participants were recruited from if different to where it was conducted
3.7 Socio-demographics	Any details e.g. education level, salary etc.
3.8 Ethnicity	No. of participants from different ethnicities and if participation was limited based on ethnicity
3.9 Additional notes	If you have any other concerns/thoughts about this section, put them here.

4. App Intervention + associated content

Item	**Explanation**
4.1 No. of groups	How many participant groups were there – including how many comparison groups or focus groups
4.2 App name	Name of app(s) being assessed. Only list apps that have the functions of interest – control apps that don't are not of interest
4.3 Size of app	Size of app – in MB
1.4 QA_Security and privacy	Provide details of any security and privacy features that app has including but not limited to: • Privacy protection • Presence and availability of security policies • Import/export practices • Use of login passwords • Encryption • Cloud back up • Adherence to Data Protection standards
4.5 Brief description of app, what it does and how	Can be copy and pasted from paper
4.6 Is app tailored/personalised	Brief description of techniques or methods used to tailor/personalise app/app content and what is used to do so e.g. names, sedentary status, stage of change, user-selected activities/PA, user-provided data of some sort etc.
4.7 QA_Was PA measured/captured within the app OR as part of the intervention (not for outcome purposes only)?	Yes/No/Unreported If physical activity was captured in a way that the user did NOT receive information about their physical activity levels at all, then code NO. For example, code NO if: • Physical activity is not captured/measured

	• Physical activity is captured for outcome purposes ONLY and the users never see the data, i.e. having access to their PA data is NOT part of the intervention at all Code YES if: • Physical activity is captured/measured using the app and users are able to see the results during the intervention • Physical activity is NOT measured/captured by the app, but is captured by some external device/questionnaire and users are given information on their physical activity levels as part of the intervention e.g. users may be provided with a separate pedometer which will show them their steps throughout the intervention period.
4.8 QA_How?	Use lists reported in Taylor 2014, pg. 48-55, and Strath et al 2013 (appendix 3 & 4 as separate documents). Provide: • name of tool/device/wearable • make and model if appropriate • settings for tool/device as necessary • self-report or objective • just the name of validated questionnaire and key citation if unaltered • alterations to validated questionnaires • full details of any unvalidated/new tool/survey/questionnaire – if multiple questions, provide location in paper rather than copying them all out. If a smartphone specific app is used, name it. If the smartphone's inbuilt accelerometer/pedometer or other sensor is used, specify this so that a distinction can be made between wearables and device-based measures.
4.9 Targeting PA and/or other behaviours?	Does the app ONLY promote PA or does it promote other health behaviours too? If multiple behaviours, list them, otherwise just say PA
4.10 Type of PA targeted and list of specific activities if specified	List if it's a specific sport/type or if its lifestyle activity, or if the activity is 'user defined' and therefore it could have been a variety of unknown activities Provide list of promoted activities if available
4.11 QA_Presence of	Specify any PA recommendations that are

PA recommendations for adults or other (specify)	provided as part of the app – if none, say no.
4.12 What type of mood was captured?	Describe the type of mood that was captured – if it appears to be PA-contingent mood i.e. mood that has been generated or is directly related to performing an activity e.g. it's being measured immediately after an activity or during an activity. Generic mood might be captured during the day without being specifically associated with an activity. Also detail if a particular type of mood was being captured (how stressed do you feel?) rather than mood in general (how do you feel?) e.g. stress levels, sadness, anxiety.
4.13 How was mood captured?	Describe how the app (or otherwise) captured participant mood including tool, units and frequency
4.14 How was mood processed and then fed back in the context of PA?	Describe how the app (or otherwise) processed (by researchers or automatically by app) and fed back participant mood in the context of PA including tool, units and frequency. If feedback on mood was provided by anything other than the app, describe how/by who and in what format and how it was delivered in the context of PA
4.15 How long was fed back mood data available for users?	Was the data available for a day, week, month, entire study duration?

4.16 QA_BCTs	Using BCT Taxonomy v1.0 (separate appendix 5), list the behaviour change techniques used in the app to the best of your ability. This includes those used to target OTHER behaviours, not just PA, in lifestyle apps. Provide:
	• An indication of the strength of your conviction regarding whether or not the BCT is present using the following '+' indicators and rules
	  ○ Present + (BCT present in all probability but evidence unclear)
	  ○ Present ++ (BCT present beyond all reasonably doubt)
	
	Authors may explicitly report the BCTs they used, but if more BCTs seem apparent then list them and note that the authors did not specify that they were part of the app.

	Please note the coding principles taken from the BCT online training: 1) Learning principle 1: Only code BCTs that are directly applied to the target behaviour(s) and population(s). *Note that for apps that are targeting multiple behaviours, ALL behaviours may be the target behaviour for a BCT.* 2) Learning principle 2: Do not infer the presence of a BCT 3) Learning principle 3: Take care distinguishing between BCTs that differ in terms of their behaviour change type (i.e. behaviour versus outcome) 4) Learning principle 4: Code technical terms and packages of BCTs that map onto BCTs in the taxonomy 5) Learning principle 5: All BCT definitions in the taxonomy include an action verb
4.17 QA_PA BCTs	Using the list of BCTs associated with PA (appendix 6 at end of this document), copy across those BCTs reported in 4.16 that match the list. If some are clearly being used to target DIFFERENT behaviours, make a note of that, but still report them.
4.18 QA_Involvement of users in development/pilot testing	Give details of if and how potential users were involved in the development/design/testing of the app.
4.19 QA_Demographics of users involved in development/pilot testing	Give demographic details of said potential users including age, sex, ethnicity if provided OR specify if different to target population/sample population listed in ‘Participants using app’ section #, above
4.20 Was anything else used along with the app, to provide the PA and mood elements of the intervention? (e.g. wearable, website, counselling session)	Some apps may not be standalone but may have been supported by other things such as websites, wearables etc. – give details
4.21 Describe how these additional tools were used	Describe how they were used for the behaviours of interest – PA and affect/mood, capturing these and providing feedback on them.
4.22 Additional notes	If you have any other concerns/thoughts about this section, put them here.

5. Outcomes

Item	**Explanation**
5.1 QA_Was usage or compliance captured/assessed?	Yes/No/Unreported Not to be confused with engagement, usage or compliance refers to responses to app content such a required step-count submissions or response to prompts to complete questions. Can include rates of usage, number of time app opened, views of a certain app pages etc.
5.2 QA_How? Including units of measurement if applicable	Provide name of tool or units of measurement. Can include rates of usage.
5.3 QA_Was app quality measured/assessed?	Yes/No/Unreported This does not include usage, captured above. But can include measures of the following as defined in appendix 7 at end of this document: • Acceptability/Participant perceptions • Aesthetics • Credible/trustworthy/appropriate or useful/essential information • Engagement • Functionality/Usability If terminology in paper is not the same as above, but definition seems to match, use paper terminology and mention which of the above definitions for the above terms you think it matches, for clarity. Or quality in general e.g. an overall quality score. Can also include general perceptions of quality for example via focus groups/piloting/interviews/user-testing/user reviews.
5.4 QA_How? Including unit of measurement if applicable	Provide name of tool if appropriate (e.g. Mobile App Rating Scale) – self-report or objective and any alterations or settings for tool as necessary. If quality assessed more generically in terms of testing, report here. If unsure of relevance of measure, add and flag as unsure.
5.5 Additional notes	If you have any other concerns/thoughts about this section, put them here.

6. Results

Item	Explanation
6.1 QA_Usage results	Results of usage assessment – as reported in 5.1/5.2
6.2 QA_Quality measure results	Results of the quality assessment measure- as reported in 5.3/5.4 Report findings both during development AND from final version of app if available and distinguish between the two.
6.3 QA_Were quality issues addressed and if so, how?	Yes/No/Unreported Report findings both during development AND from final version of app if available and distinguish between the two. Describe briefly what changes were made based results from quality assessment, this could include changes to content/availability/functionality etc. based on feedback from users or poor engagement etc.
6.4 Additional notes	If you have any other concerns/thoughts about this section, put them here.

7. Other

Item	Explanation
7.1 Any other comments/discussion/ conclusions relevant to review questions	Note any other passages in the paper that you think are relevant to the review questions that haven't been captured elsewhere. **Research questions:** The purpose of this data extraction form is to collect data from eligible literature to answer the following research questions: 1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change? 2. What are the characteristics and content of physical activity apps that include feedback on immediate affect? 3. What is the quality of these physical activity apps, that include feedback on immediate affect?
7.2 References to other relevant studies/linked papers	Note references that may provide more information on the app or another potentially relevant app

reporting on app further	

B: QA_TCS: Theoretical Coding Scheme

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

TCS generic guidelines, adapted for thesis

The coding scheme that follows comprises 19 items, only 11 of which are captured as part of this data extraction/quality assessment. For each item, code as Yes or No. When 'Yes', state the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).

The coding scheme is based on the *explicit* use of theory. Consequently, even when theory-relevant information is implied but is not stated explicitly, the related items should be coded as 'no'.

<u>**Defining Terms**</u>

Please refer to the definitions (see below) during coding:

Theory (or Model)
'a set of interrelated concepts, definitions and propositions that present a *systematic* view of events or situations by specifying relations among variables, in order to *explain* or *predict* the events or situations' (Glanz & Rimer, 1995). Examples include: Theory of Planned Behaviour (TPB), Theory of Reasoned Action (TRA), Health Belief Model (HBM), stage of change/trans-theoretical model etc. Please refer to the '<u>*Table of Theories*</u>' below, but if the theory or construct is not represented in this table, refer to 'ABC of behaviour change theories' book (separate appendix 8). Try searching for the construct to see which theories it might relate to.

Theory-relevant construct
A construct (a key concept, excluding behaviour) within a theory/model upon which the intervention is based. Please refer to the '<u>*Table of Theories*</u>' below to identify whether a particular construct belongs to the specified theory. If the theory or construct is not represented in this table, refer to 'ABC of behaviour change theories' book (separate appendix 8). Try searching for the construct to see which theories it might relate to.

Predictors
Constructs that are not explicitly linked to a theory by the authors, but are targeted for intervention (as a means to change behaviour) because they predict behaviour.

Predictors must only be coded if the author has presented evidence that the construct predicts/correlates with/causes behaviour. Predictors do not include actual behaviour, self-reported or otherwise (e.g. amount of time spent exercising), or biological factors (e.g., age, sex).

Intervention Technique
Strategy used to change behaviour, theory-relevant construct or predictor (e.g., providing information on consequences; prompting specific goal setting; prompting barrier identification; modelling the behaviour; planning social support). **NOTE: The TCS was developed before the first version of the BCT taxonomy was released, but BCTs can be coded in this section**.

Table of Theories

Theory	***Theory-relevant constructs***
Theory of Reasoned Action	1. **Attitudes**; 2. **Subjective Norms**; 3. **Intentions**
Theory of Planned Behaviour	1. **Attitudes**; 2. **Subjective Norms**; 3. **Perceived Behavioural Control**; 4. **Intentions**
Trans-Theoretical Model/Stages of Change	1. **Self-Efficacy** (person's confidence in performing a particular behaviour); 2. **Decisional Balance**
Social Cognitive Theory	1. **Self-Efficacy** (person's confidence in performing a particular behavior)/ **Behavioral capability** (Knowledge and skill to perform a given behavior); 2. **Action-Outcome Expectancies** (extent that one's actions are seen as instrumental for the outcome/values associated with outcomes)/**attitudes**; 3. **Barriers** (including changes to environment/emotional barriers or one's perceptions of them). The following constructs are also related to the theory (and subsequently should be listed when they are cited by the authors): *Behavioral capability:* Knowledge and skill to perform a given behavior; *Attitudes (outcome-expectancies):-* *Expectations:* Anticipatory outcomes of a behavior; *Expectancies:* The values that the person places on a given outcome, incentives; *Self-control:* Personal regulation of goal-directed behavior or performance

	Goals?
Health Belief Model	1. **Perceived Susceptibility**; 2. **Perceived Severity**; 3. **Perceived Benefits** (one's belief in the efficacy of the advised action to reduce risk or seriousness of impact); 4. **Perceived Barriers**; 5. **Cues to Action**; 6. **Self-Efficacy** (added by Rosenstock et al., 1988)
Protection Motivation Theory (PMT)	1. **Intention**; 2. **Perceived Behavioural Control**; 3. **Perceived Severity**; 4. **Perceived Vulnerability**; 5. **Response Efficacy**; 6. **Response Costs**; 7. **Fear** (now added to the model),
Rubicon Model / Model of Action Phases (Heckhausen, 1991; Gollwitzer, 1990)	1. **Motivation** (e.g., intention); 2. **Volition** (e.g., planning)

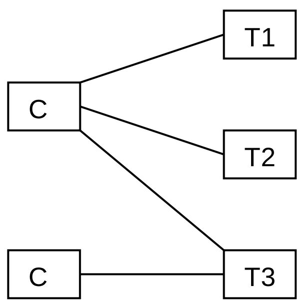
Notes:

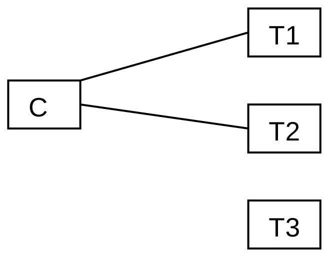
1. When more than one intervention is used (excluding the control group), the items should be coded separately for each intervention
2. Note for Stage of Change/Trans-Theoretical Model: While the construct 'stage of change' may either be used to select recipients of an intervention (item 4) or to tailor an intervention (item 6), 'stage of change' measures should not be coded for items 7-11. However, self-efficacy and decisional balance, both constructs within the TTM, should be coded throughout.
3. Where multiple behaviours are targeted by an app, if different theories/constructs/techniques are used to target different behaviours, report ALL and distinguish between those for different behaviours if possible.

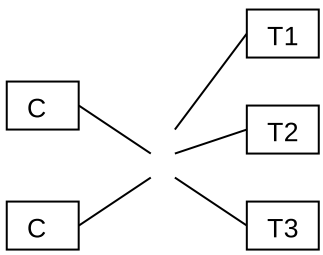
8. TCS item specific coding descriptions

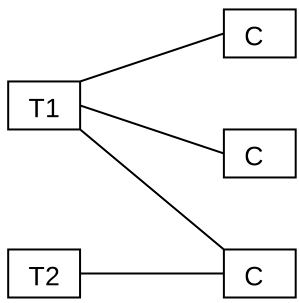
Item	**Explanation**
8.1a Theory/model of behaviour mentioned	Using the drop down menu, enter: Yes/No/Don't know Models/theories that specify relations among variables, in order to *explain* or *predict* behaviour (e.g., TPB, SCT, and HBM) are mentioned, even if the intervention is not based on this theory.
8.1b Theories and paper location	List the theory/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).

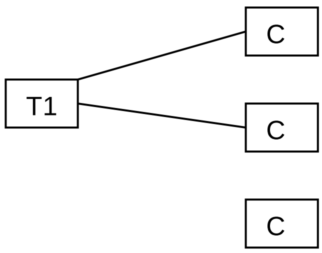
8.2a Targeted construct mentioned as predictor of behaviour	Using the drop down menu, enter: Yes/No/Don't know 'Targeted' construct refers to a psychological construct that the study intervention is hypothesised to change. Evidence that the psychological construct relates to (correlates/predicts/causes) behaviour should be presented within the introduction or method (rather than the Discussion).
8.2b Construct and location of evidence that construct relates to behaviour	List the construct/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]). (These details are often found in the introduction section.)
8.2c Location that this predictor is targeted by the intervention	List the predictor/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]). (These details are often found in the methods section.)
8.3a Intervention based on single theory?	Using the drop down menu, enter: Yes/No/Don't know The intervention is based on a single theory (rather than a combination of theories or theory + predictors)
8.3b Location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.4a Theory/ predictors used to select recipients for the intervention	Using the drop down menu, enter: Yes/No/Don't know Participants were screened/selected based on achieving a particular score/level on a theory-relevant construct/predictor
8.4b Construct (theory) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.4c Predictor and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.5a Theory/ predictors used to select/develop intervention techniques	Using the drop down menu, enter: Yes/No/Don't know The intervention is explicitly based on a theory or predictor or combination of theories or predictors (These details are often found in the Methods section and may be a replication of the findings for

	item 8.2.)
8.5bTheory and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.5c Predictor and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.6a Theory/ predictors used to tailor intervention techniques to recipients	Using the drop down menu, enter: Yes/No/Don't know The intervention differs for different sub-groups that vary on a psychological construct (e.g., stage of change) or predictor at baseline
8.6b Construct and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.6c Predictor and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.7a All intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor	Using the drop down menu, enter: Yes/No/Don't know Each intervention technique (T) is linked to at least one theory-relevant construct/predictor (C). Examples of meeting this criterion: - T1 & T2 are used to change C1; T3 is used to change C1 and C2. Techniques 1-3 are the only intervention techniques used (see DIAGRAM 1). - T1 used to change C1; T2 used to change C2; T3 used to change C3 **DIAGRAM 1: Example: All intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)**  In some cases, items 8.7, 8.8 and 8.9 may all be coded as 'no' (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or

	vice-versa). If item 8.7 is coded ‘yes’ then items 8.8 and 8.9 mus coded ‘no’. In instances where the name of the *construct* might overlap with the name of the technique (e.g., tips on *social support* were given; exercise *benefits* were discussed), code this as a direct link between technique and construct (i.e. they do not need to be written in the form ‘tips on social support [technique] were used to target social support [construct]’; exercise benefits were discussed [technique] to target perceived benefits [construct]).
8.7b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.7c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.8a At least one, but not all, of the intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor	Using the drop down menu, enter: Yes/No/Don’t know At least one, but not all, of the intervention techniques (T) are explicitly linked to at least one theory-relevant construct/predictor (C). At least one technique, therefore, is not linked to any theory-relevant construct/predictor. Examples of meeting this criterion: - T1 and T2 are used to change C1; T3 is used but not linked to a construct/predictor (see DIAGRAM 2). - T1 used to change C1; T2 used to change C1 and C2; T3 is used but not linked to a construct/predictor. **DIAGRAM 2: Example: At least one, but not all, of the intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)** 

	In some cases, items 8.7, 8.8 and 8.9 may all be coded as ‘no’ (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa). If item 8.7 is coded ‘yes’ then items 8.8 and 8.9 must be coded ‘no’. It is possible for items 8.8 and 8.9 to both be coded as ‘yes’.
8.8b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.8c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.9a Group of techniques are linked to a group of constructs/ predictors	Using the drop down menu, enter: Yes/No/Don’t know A cluster/group of techniques (T) is linked to a cluster/group of theory-relevant constructs/ predictors (C). Example of meeting this criterion: - T1, T2, T3 are used to change C1, C2 (see DIAGRAM 3). **DIAGRAM 3: Example: Group of techniques (T) are linked to a group of theory-relevant constructs/predictors (C)**  In some cases, items 8.7, 8.8 and 8.9 may all be coded as ‘no’ (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa). If item 8.7 is coded ‘yes’ then items 8.8 and 8.9 must be coded ‘no’. It is possible for items 8.8 and 8.9 to both be coded as ‘yes’.
8.9b List clusters of	List the supporting evidence and its location, as

techniques/constructs and location	follows: ([insert page number], [insert paragraph number]).
8.9c List clusters of techniques/predictors and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.10a All theory-relevant constructs/predictors are explicitly linked to at least one intervention technique	Using the drop down menu, enter: Yes/No/Don’t know <u>Every</u> theoretical construct within a stated theory, or every stated predictor (C) (see item 8.5), is linked to at least one intervention technique (T) Examples of meeting this criterion: - C1 & C2 are linked to T1; C3 is linked to T2. Constructs 1-3 are the only constructs within the theory specified in item 5 (see DIAGRAM 4). - C1 is linked to T1; C2 is linked to T2; C3 is linked to T3 To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the ‘*Table of Theories*’. **DIAGRAM 4: Example: All constructs within a stated theory/all predictors (C) are linked to at least one intervention technique (T)** 
8.10b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.10c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.11a At least one, but not all, of the theory relevant constructs/predictors are explicitly linked to at least one intervention technique	Using the drop down menu, enter: Yes/No/Don’t know At least one, but not all, of the theoretical constructs within a stated theory or at least one, but not all, of the stated predictors (see item 8.5) are linked to at least one intervention technique.

	At least one, but not all, of the constructs within a stated theory/all stated predictors (C) (see item 5) are linked to at least one intervention technique (T). At least one construct within a stated theory, therefore, is not linked to any intervention technique. Examples of meeting this criterion: - C1 and C2 are linked to T1; C3 is not linked to an intervention technique- or is not highlighted by the authors (see DIAGRAM 5). - C1 is linked to T1; C2 is linked to T1 and T2; C3 is not linked to an intervention technique- or is not highlighted by the authors. To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the '*Table of Theories*'. **DIAGRAM 5: Example: At least one, but not all, of the constructs within a stated theory/stated predictors (C) are linked to at least one intervention technique (T)** 
8.11b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.11c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).
8.12 Additional notes	If you have any other concerns/thoughts about this section, put them here.

C: QA_ PA BCTs: Behaviour Change Techniques with an evidence base for promoting/inhibiting physical activity promotion

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

9. PA BCTS

Using the data entered in QA_PA BCTs in the Data extraction tab, complete as follows:

- 9.1 Where a ‘Good BCT for PA’ has been used, place a 1 in the corresponding cell, UNLESS it’s one of the two following BCTs, in which case, put a 2 in the corresponding cell:
    - Behaviour practice/rehearsal
    - Prompts/cues
- 9.2 Where a ‘BCT with mixed evidence for PA’ has been used, place a 1 in the corresponding cell
- 9.3 Where a ‘Bad BCT for PA’ has been used, place a -1 in the corresponding cell

Totals for each set of BCTs will be automatically calculated. Do not enter any data into BCT cells that are not applicable as this will make the automatic calculation stop working.

D: QA_MARS: The Mobile App Rating Scale

It is important that you have downloaded the app, if available, at this point! If the app is not available for download, complete the following items to the best of your ability using the information provided in the paper.

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

10. App admin

Item	Explanation
10.1 App name	List the name of the relevant app If multiple relevant apps are reported, use a new line in the spreadsheet for each
10.2 Version (paper)	List the version number of the app according to the paper
10.3 Version (sourced)	If the version reported in the paper cannot be sourced for full data extraction and testing (download and 2 day use), specify the version that was sourced
10.4 Rating (sourced)	Provide the average star rating for the app that has been sourced for full data extraction and testing (download and 2 day use), if available i.e. if it’s been downloaded from an app store. If a star rating is provided in the paper, report this

	too and the version of the app it coincides with.
10.5 Developer (if not authors)	Provide details of the developers of the app – this might be in the paper, or, if the app is in the app store, it might be reported there (e.g. in Google Play this is at the bottom of the app page) or in the details of the app itself, e.g. try going to the 'About' menu item in the app.
10.6 Relevant expertise/credentials of development team (NON-MARS ITEM)	Provide any information available on the credentials or relevant expertise of the development team such as: • Behaviour change • Medicine • App development/programming • Physical activity/exercise
10.7 Number of ratings for this version	If the app is available in an app store, capture the number of ratings that have been provided for it. For example, in Google Play, this number will be in the top right corner of the app page next to a person icon and the average star rating. If the number of ratings is provided in the paper, report this too and the version of the app it coincides with.
10.8 QA_Release date (NON-MARS ITEM)	If reported, provide the release date for the app (most likely reported in app store link) otherwise enter NA
10.9 Last update	If the app is available in an app store, capture the date it was last updated. For example, in Google Play, this is at the bottom of the app page. If the last update is provided in the paper, report this too and the version of the app it coincides with.
10.10 Cost	If the app is available in an app store, capture the cost or if it was free to download. If the cost is provided in the paper, report this too and the version of the app it coincides with.
10.11 Platform	Report the operating system (platform) that the app is compatible with. E.g. Android, iOS (Apple) or another platform such as Windows, Blackberry and so on. If you are unsure, put what you think it is but also say 'unsure'.
10.12 Affiliations	Enter all that apply: • Unknown • Commercial • Government • NGO • University

	Provide evidence for this and/or specify where this is reported in the paper/app store.
10.13 Technical aspects	Enter all that apply: • Allows sharing (Facebook, Twitter, etc.) - specify • Has an app community (sharing with other users only) • Allows password-protection • Requires login • Sends reminders • Needs web access to function If app cannot be accessed for download, use any information available in paper, otherwise enter 'not reported'
10.14 Additional notes	If you have any other concerns/thoughts about this section, put them here.

FOR THE FOLLOWING SECTIONS, BASE YOUR ANSWERS ON THE FINAL VERSION OF THE APP. Do not use developmental information to complete this section unless it refers to the final iteration of the app described in the paper, i.e. no further changes were made to the app following feedback/comments/piloting of the app. If unsure, ask LL before you start to code.

FOR APPS TARGETING MULTIPLE BEHAVIOURS USE INFORMATION ABOUT THE WHOLE APP, NOT JUST THOSE BIT RELATED TO CHANGING PA

The Rating scale assesses app quality on four dimensions. All items are rated on a 5-point scale from "1.Inadequate" to "5.Excellent" **but some allow you to answer NA too**. Enter the number that most accurately represents the quality of the app component you are rating. Please use the descriptors provided for each response category.

IF YOU ARE UNABLE TO COMPLETE A CODE DUE TO NO INFORMATION BEING REPORTED, ENTER 'NOT REPORTED'

11. Engagement

Engagement – fun, interesting, customisable, interactive (e.g. sends alerts, messages, reminders, feedback, enables sharing), well-targeted to audience

Item	**Explanation**
11.1a Entertainment: Is the app fun/entertaining to use?	Enter ONE of the numbers below that best corresponds to your perception of the app:

Does it use any strategies to increase engagement through entertainment (e.g. through gamification)? Explain answer	1. Dull, not fun or entertaining at all 2. Mostly boring 3. OK, fun enough to entertain user for a brief time (< 5 minutes) 4. Moderately fun and entertaining, would entertain user for some time (5-10 minutes total) 5. Highly entertaining and fun, would stimulate repeat use If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
11.1b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
11.2a Interest: Is the app interesting to use? Does it use any strategies to increase engagement by presenting its content in an interesting way?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Not interesting at all 2. Mostly uninteresting 3. OK, neither interesting nor uninteresting; would engage user for a brief time (< 5 minutes) 4. Moderately interesting; would engage user for some time (5-10 minutes total) 5. Very interesting, would engage user in repeat use

	If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
11.2b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
11.3a Customisation: Does it provide/retain all necessary settings/preferences for apps features (e.g. sound, content, notifications, etc.)?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Does not allow any customisation or requires setting to be input every time 2. Allows insufficient customisation limiting functions 3. Allows basic customisation to function adequately 4. Allows numerous options for customisation 5. Allows complete tailoring to the individual's characteristics/preferences, retains all settings If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from

	users/testers.
11.3b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
11.4a Interactivity: Does it allow user input, provide feedback, contain prompts (reminders, sharing options, notifications, etc.)? Note: these functions need to be customisable and not overwhelming in order to be perfect.	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No interactive features and/or no response to user interaction 2. Insufficient interactivity, or feedback, or user input options, limiting functions 3. Basic interactive features to function adequately 4. Offers a variety of interactive features/feedback/user input options 5. Very high level of responsiveness through interactive features/feedback/user input options If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
11.4b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted?

	Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
11.5a Target group: Is the app content (visual information, language, design) appropriate for your target audience?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Completely inappropriate/unclear/confusing 2. Mostly inappropriate/unclear/confusing 3. Acceptable but not targeted. May be inappropriate/unclear/confusing 4. Well-targeted, with negligible issues 5. Perfectly targeted, no issues found If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
11.5b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
11.6 Engagement mean score	Do not enter anything here

12. Functionality

Functionality – app functioning, easy to learn, navigation, flow logic, and gestural design of app

Item	Explanation
12.1 Performance: How accurately/fast do the app features (functions) and components (buttons/menus) work?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. App is broken; no/insufficient/inaccurate response (e.g. crashes/bugs/broken features, etc.) 2. Some functions work, but lagging or contains major technical problems 3. App works overall. Some technical problems need fixing/Slow at times 4. Mostly functional with minor/negligible problems 5. Perfect/timely response; no technical bugs found/contains a 'loading time left' indicator If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
12.2 Ease of use: How easy is it to learn how to use the app; how clear are the menu labels/icons and instructions?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No/limited instructions; menu labels/icons are confusing; complicated 2. Useable after a lot of time/effort 3. Useable after some time/effort 4. Easy to learn how to use the app (or has clear instructions) 5. Able to use app immediately; intuitive; simple If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers.

	Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
12.3 Navigation: Is moving between screens logical/accurate/appropriate/ uninterrupted; are all necessary screen links present?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Different sections within the app seem logically disconnected and random/confusing/navigation is difficult 2. Usable after a lot of time/effort 3. Usable after some time/effort 4. Easy to use or missing a negligible link 5. Perfectly logical, easy, clear and intuitive screen flow throughout, or offers shortcuts If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
12.4 Gestural design: Are interactions (taps/swipes/pinches/scrolls) consistent and intuitive across all components/screens?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Completely inconsistent/confusing 2. Often inconsistent/confusing 3. OK with some inconsistencies/confusing elements 4. Mostly consistent/intuitive with negligible problems 5. Perfectly consistent and intuitive If the app is not available for download, use the information in the paper to select a SINGLE number and explain why, data can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.

12.5 Functionality mean score	Do not enter anything here

13. Aesthetics

Aesthetics – graphic design, overall visual appeal, colour scheme, and stylistic consistency

Item	Explanation
13.1 Layout: Is arrangement and size of buttons/icons/menus/content on the screen appropriate or zoomable if needed?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Very bad design, cluttered, some options impossible to select/locate/see/read. Device display not optimised 2. Bad design, random, unclear, some options difficult to select/locate/see/read 3. Satisfactory, few problems with selecting/locating/seeing/reading items or with minor screen-size problems 4. Mostly clear, able to select/locate/see/read items 5. Professional, simple, clear, orderly, logically organised, device display optimised. Every design component has a purpose If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
13.2 Graphics: How high is the quality/resolution of graphics used for buttons/icons/menus/content?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Graphics appear amateur, very poor visual design - disproportionate, completely stylistically inconsistent 2. Low quality/low resolution graphics; low quality visual design – disproportionate, stylistically

	inconsistent 3. Moderate quality graphics and visual design (generally consistent in style) 4. High quality/resolution graphics and visual design – mostly proportionate, stylistically consistent 5. Very high quality/resolution graphics and visual design - proportionate, stylistically consistent throughout If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
13.3 Visual appeal: How good does the app look?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No visual appeal, unpleasant to look at, poorly designed, clashing/mismatched colours 2. Little visual appeal – poorly designed, bad use of colour, visually boring 3. Some visual appeal – average, neither pleasant, nor unpleasant 4. High level of visual appeal – seamless graphics – consistent and professionally designed 5. As above + very attractive, memorable, stands out; use of colour enhances app features/menus If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
13.4 Aesthetics mean score	Do not enter anything here

14. Information

Information – Contains high quality information (e.g. text, feedback, measures, and references) from a credible source. Select N/A if the app component is irrelevant.

Item	**Explanation**
14.1 Accuracy of app description (in app store): Does app contain what is described?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Misleading. App does not contain the described components/functions. Or has no description 2. Inaccurate. App contains very few of the described components/functions 3. OK. App contains some of the described components/functions 4. Accurate. App contains most of the described components/functions 5. Highly accurate description of the app components/functions If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers. IF APP DESCRIPTION NOT PROVIDED ENTER 'NOT PROVIDED'
14.2a Goals: Does app have specific, measurable and achievable goals (specified in app store description or within the app itself)?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A Description does not list goals, or app goals are irrelevant to research goal (e.g. using a game for educational purposes) 1. App has no chance of achieving its stated goals 2. Description lists some goals, but app has very little chance of achieving them 3. OK. App has clear goals, which may be achievable. 4. App has clearly specified goals, which are measurable and achievable

	5. App has specific and measurable goals, which are highly likely to be achieved If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
14.2b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
14.3a Quality of information: Is app content correct, well written, and relevant to the goal/topic of the app?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A There is no information within the app 1. Irrelevant/inappropriate/incoherent/incorrect 2. Poor. Barely relevant/appropriate/coherent/may be incorrect 3. Moderately relevant/appropriate/coherent/and appears correct 4. Relevant/appropriate/coherent/correct 5. Highly relevant, appropriate, coherent, and correct If the app is not available for download, use the information in the paper to select a SINGLE

	number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
14.3b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
14.4a Quantity of information: Is the extent coverage within the scope of the app; and comprehensive but concise?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A There is no information within the app 1. Minimal or overwhelming 2. Insufficient or possibly overwhelming 3. OK but not comprehensive or concise 4. Offers a broad range of information, has some gaps or unnecessary detail; or has no links to more information and resources 5. Comprehensive and concise; contains links to more information and resources If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments

	from users/testers.
14.4b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
14.5a Visual information: Is visual explanation of concepts – through charts/graphs/images/videos, etc. – clear, logical, correct?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A There is no visual information within the app (e.g. it only contains audio, or text) 1. Completely unclear/confusing/wrong or necessary but missing 2. Mostly unclear/confusing/wrong 3. OK but often unclear/confusing/wrong 4. Mostly clear/logical/correct with negligible issues 5. Perfectly clear/logical/correct If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
14.5b If your app target multiple behaviours, what promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted?

	Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
14.6 Credibility: Does the app come from a legitimate source (specified in app store description or within the app itself)?	For this item, you may need to go online to check the legitimacy of the source. Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Source identified but legitimacy/trustworthiness of source is questionable (e.g. commercial business with vested interest) 2. Appears to come from a legitimate source, but it cannot be verified (e.g. has no webpage) 3. Developed by small NGO/institution (hospital/centre, etc.) /specialised commercial business, funding body 4. Developed by government, university or as above but larger in scale 5. Developed using nationally competitive government or research funding (e.g. Australian Research Council, NHMRC) If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
14.7 Evidence base: Has the app been trialled/tested; must be verified by evidence (in published scientific literature)?	For this item, you may have been provided with additional papers, if not, base your answer on the information provided in the original paper. Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app:

	N/A The app has not been trialled/tested 1. The evidence suggests the app does not work 2. App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has partially positive outcomes in studies that are not randomised controlled trials (RCTs), or there is little or no contradictory evidence. 3. App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has positive outcomes in studies that are not RCTs, and there is no contradictory evidence. 4. App has been trialled and outcome tested in 1-2 RCTs indicating positive results 5. App has been trialled and outcome tested in > 3 high quality RCTs indicating positive results If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
14.8 Information mean score	Do not enter anything here.

15. App quality mean score

This will be automatically calculated, do not enter anything here.

(NOTE TO LL: * Exclude questions rated as “N/A” from the mean score calculation.)

16. Perceived Impact

These added items can be adjusted and used to assess the perceived impact of the app on the user’s knowledge, attitudes, and intentions to change as well as the likelihood of actual change in the target health behaviour.

Item	Explanation
16.1 Awareness: This app is likely to increase awareness of the importance of addressing physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
16.2 Knowledge: This app is likely to increase knowledge/understanding of physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
16.3 Attitudes: This app is likely to change attitudes toward improving physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE

	number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
16.4 Intention to change: This app is likely to increase intentions/motivation to address physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
16.5 Help seeking: Use of this app is likely to encourage further help seeking for physical activity (if it's required)	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.
16.6 Behaviour change: Use of this app is likely increase physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree

	2. 3. 4. 5. Strongly agree If the app is not available for download, use the information in the paper to select a SINGLE number. Evidence can come from descriptions, screenshots or relevant feedback/comments from users/testers. Provide evidence for your coding, if app not available, data can come from descriptions, screenshots or relevant feedback/comments from users/testers.

E: QA_PA measure reliability: Reliability and validity of the measures of physical activity

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

4.6/4.7 PA measure

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

17. Evaluation of PA measure used to capture and feedback PA to users

Item	Explanation
17.1 Evidence for/against measure?	If paper reports any evidence for/against the measure, report it here. This includes any measures of reliability or validity that may be reported. If paper cites other papers that provide evidence for or against the measurement tool, add them here for LL follow up. Most often cited in Methods or Discussion.
17.2 Strengths of measure	Provide details of any strengths of the measure reported in the paper, or any you perceive. For example any preferences users had for the measure, ease of use, brevity, perceived accuracy etc. If it's a self-report measure ALSO enter: 'self-report

	measure strengths' If it's an objective measure ALSO enter: 'objective measure strengths'
17.3 Weaknesses of measure	Provide details of any weaknesses of the measure reported in the paper, or any you perceive. For example any difficulties the users had with the measure, functional problems or misunderstandings or lack of wear/completion etc. If it's a self-report measure ALSO enter: 'self-report measure weaknesses' If it's an objective measure ALSO enter: 'objective measure weaknesses'
17.4 Additional notes	If you have any other concerns/thoughts about this section, put them here.

Appendices

NB. Appendices 2, 3, 4, 5 and 8 are not attached as they refer to separate book chapters/books/papers that can be accessed

Appendix 1 – Tables of characteristics of study designs x 3 (below)

Appendix 2 –Nieswiadomy, R.M. 2012. Chapter 10: Qualitative research designs. *Foundations in nursing research*., pp.171–184.

Appendix 3 – Taylor, N. 2014. Challenges in measuring physical activity in the context of mental health *In*: A. Clow and S. Edmunds, eds. *Physical Activity and Mental Health*. Leeds: Human Kinetics, pp.41–61.

Appendix 4 - Strath, S.J., Kaminsky, L.A., Ainsworth, B.E., Ekelund, U., Freedson, P.S., Gary, R.A., Richardson, C.R., Smith, D.T. and Swartz, A.M. 2013. Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the American Heart association. *Circulation*. **128**(20), pp.2259–2279.

Appendix 5 – Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M.P., Cane, J. and Wood, C.E. 2013. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine*. [Online]. **46**(1), pp.81–95. Available from: http://link.springer.com/10.1007/s12160-013-9486-6.

Appendix 6 – PA BCTs (below)

Appendix 7 – Quality indicators and definitions (below)

Appendix 8 - Michie, S., West, R., Campbell, R., Brown, J. and Gainforth, H. 2014. *ABC of behaviour change theories: an essential resource for researchers, policy makers and practitioners* [Online]. Silverback Publishing. Available from: https://www.waterstones.com/book/abc-of-behaviour-change-theories/susan-michie/robert-west/9781912141012.

Appendix 1 - List of study design features (studies with allocation to interventions at the individual level)

Courtesy of Cochrane Handbook

	RCT	Q-RCT	NRCT	CBA	PCS	RCS	HCT	NCC	CC	XS	BA	CR/CS
Was there a comparison:												
Between two or more groups of participants receiving different interventions?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Within the same group of participants over time?	P	P	N	Y	N	N	N	N	N	N	Y	N
Were participants allocated to groups by:												
Concealed randomization?	Y	N	N	N	N	N	N	N	N	N	na	na
Quasi-randomization?	N	Y	N	N	N	N	N	N	N	N	na	na
By other action of researchers?	N	N	Y	P	N	N	N	N	N	N	na	na
Time differences?	N	N	N	N	N	N	Y	N	N	N	na	na
Location differences?	N	N	P	P	P	P	P	na	na	na	na	na
Treatment decisions?	N	N	N	P	P	P	N	N	N	P	na	na
Participants' preferences?	N	N	N	P	P	P	N	N	N	P	na	na
On the basis of outcome?	N	N	N	N	N	N	N	Y	Y	P	na	na
Some other process? (specify)												
Which parts of the study were prospective:												
Identification of participants?	Y	Y	Y	P	Y	N	P*	Y	N	N	P	P
Assessment of baseline and allocation to intervention?	Y	Y	Y	P	Y	N	P*	Y	N	N	na	na
Assessment of outcomes?	Y	Y	Y	P	Y	P	P	Y	N	N	P	P
Generation of hypotheses?	Y	Y	Y	Y	Y	Y	Y	Y	P	P	P	na
On what variables was comparability between groups assessed:												
Potential confounders?	P	P	P	P	P	P	P	P	P	P	N	na
Baseline assessment of outcome variables?	P	P	P	Y	P	P	P	N	N	N	N	na

Y=Yes; P=Possibly; P*=Possible for one group only; N=No; na=not applicable. NB: Note that 'possibly' is used in the table to indicate cells where *either* 'Y' or 'N' may be the case. It should not be used as a response option when applying the checklist; if uncertain, the response should be 'can't tell'

RCT=Randomized controlled trial; Q-RCT=Quasi-randomized controlled trial; NRCT=Non-randomized controlled trial; CBA=Controlled before-and-after study; PCS=Prospective cohort study; RCS=Retrospective cohort study; HCT=historically controlled trial; NCC=Nested case-control study; CC=Case-control study; XS=Cross-sectional study; BA=Before-and-after comparison; CR/CS=Case report/Case series.

List of study design features (studies with allocation to interventions at the group level)

Courtesy of Cochrane Handbook

	ClRCT	ClQ-RCT	ClNRT	CITS	CChBA	ITS	ChBA	EcoXS
Was there a comparison:								
Between two or more groups of clusters receiving different interventions?	Y	Y	Y	Y	Y	N	N	Y
Within the same group of clusters over time?	P	P	N	Y	N	Y	Y	N
Were clusters allocated to groups by:								
Concealed randomization?	Y	N	N	N	N	N	N	N
Quasi-randomization?	N	Y	N	N	N	N	N	N
By other action of researchers?	N	N	Y	P	P	N	N	N
Time differences?	N	N	N	Y	Y	Y	Y	N
Location differences?	N	N	P	P	P	N	N	P
Policy/public health decisions?	Na	na	P	P	P	P	na	na
Cluster preferences?	Na	na	P	P	P	P	na	na
Some other process? (specify)								
Which parts of the study were prospective:								
Identification of participating clusters?	Y	Y	Y	P	P	P	P	N
Assessment of baseline and allocation to intervention?	Y	Y	Y	P	P	P	P	N
Assessment of outcomes?	Y	Y	Y	P	P	P	P	N
Generation of hypotheses?	Y	Y	Y	Y	Y	Y	Y	P
On what variables was comparability between groups assessed:								
Potential confounders?	P	P	P	P	P	P	P	P
Baseline assessment of outcome variables?	P	P	P	Y	Y	Y	Y	N

Note that 'cluster' refers to an entity (e.g. an organization), not necessarily to a group of participants; 'group' refers to one or more clusters.

Note that ‘possibly’ is used in the table to indicate cells where *either* ‘Y’ or ‘N’ may be the case. It should not be used as a response option when applying the checklist; if uncertain, ‘can’t tell’ should be used.

Y=Yes; P=Possibly; P*=Possible for one group only; N=No; NR=Not required.

ClRCT=Cluster randomized controlled trial; ClQ-RCT=Cluster quasi-randomized controlled trial; ClNRT=Cluster non-randomized controlled trial; CITS=Controlled interrupted time series (Shadish 2002); CChBA=Controlled cohort before-and-after study; ITS=Interrupted time series; ChBA=Cohort before and after study (Shadish 2002); EcoXS=Ecological cross-sectional study.

Some types of NRS design used for evaluating the effects of interventions

Courtesy of Cochrane Handbook

Designs are distinguished below by labels in common use and descriptions are intentionally non-specific because the labels are interpreted in different ways with respect to details. The NRSMG does not advocate using these labels for reasons explained in Section 13.5.1 of Handbook.

Non-randomized controlled trial.	An experimental study in which people are allocated to different interventions using methods that are not random.
Controlled before-and-after study.	A study in which observations are made before and after the implementation of an intervention, both in a group that receives the intervention and in a control group that does not.
Interrupted-time-series study.	A study that uses observations at multiple time points before and after an intervention (the 'interruption'). The design attempts to detect whether the intervention has had an effect significantly greater than any underlying trend over time.
Historically controlled study.	A study that compares a group of participants receiving an intervention with a similar group from the past who did not.
Cohort study.	A study in which a defined group of people (the cohort) is followed over time, to examine associations between different interventions received and subsequent outcomes. A 'prospective' cohort study recruits participants before any intervention and follows them into the future. A 'retrospective' cohort study identifies subjects from past records describing the interventions received and follows them from the time of those records.
Case-control study.	A study that compares people with a specific outcome of interest ('cases') with people from the same source population but without that outcome ('controls'), to examine the association between the outcome and prior exposure (e.g. having an intervention). This design is particularly useful when the outcome is rare.
Cross-sectional study.	A study that collects information on interventions (past or present) and current health outcomes, i.e. restricted to health states, for a group of people at a particular point in time, to examine associations between the outcomes and exposure to interventions.
Case series (uncontrolled longitudinal study).	Observations are made on a series of individuals, usually all receiving the same intervention, before and after an intervention but with no control group.

Appendix 6 – BCTs associated with changes in physical activity

Positive	Negative	Mixed evidence
Biofeedback	Information about antecedents	Problem solving
Demonstration of behaviour		Reviewing behavioural goal
Behaviour Practice/rehearsal		Feedback on behaviour
Graded tasks		
Action planning		
Instruction on how to perform behaviour		
Prompts/cues		
Self-reward		
Self-monitoring of behaviour		
Providing info on health consequences		
Behaviour substitutions		
Restructuring physical environment		
Providing practical social support		
Restructuring social environment		
Unspecified social support		
Setting outcome goals		
Pros and cons		
Social rewards		
Habit formation		
Commitment		
Discrepancy between current behaviour and goals		

Appendix 7 – Quality indicators and definitions

Quality indicator		Summary definition
1	Acceptability or Participant perceptions	Positive and negative feedback or recommendations from users on the app content e.g. preferences or recommendations for information, tone or features. Can include user ratings or reviews of the app (the latter may overlap with other indicators such as 3, or 4, these should be coded separately). May include barriers and facilitators (feasibility) to use of the app and/or smartphone such as it being easy to fit self-report requests into your routine, or forgetting to carry the phone. May also include whether or not users/providers/practitioners would recommend the app to others/patients
2	Aesthetics	Visual attractiveness of the app interface design in terms of colours, fonts, and layout. How professional the design is. How pleasing to eye the design and layout is. Can include relevance of design to the behaviour.
3	Credible, trustworthy/appropriate or useful/essential information	Content of the app is likely to be accurate or believable – not making impossible or implausible claims. Content is safe for users, won't harm them or will minimise harm or provides a caveat for medical information that requires seeing a professional. Information/app appears useful.
4	Engagement	Use of methods to encourage user interactivity with the app, can include use of certain strategies or features that promote/inhibit for example, feedback, tailoring, prompts/reminders, gamification. (Often this term has also been used to refer to usage/response to app intervention features or feasibility, such as required step-count submissions or required message responses (Monroe et al., 2015). Or it has encompassed both these and the above summary definition (Rose et al., 2017). These items have been separated out here).
5	Functionality/Usability	Ease of use of the app and/or smartphone features, such as navigation, terminology, design in relation to ease of use, not aesthetics (see 4) as well as general perception of how

		much support might be required for use or how complex or inconsistent it might be. Functional errors related to app operations such as bugs/crashing also captured here. Includes practicality of use for promoting or capturing physical activity based on functions and features. Can be assessed by questionnaires such as the System Usability Scale, interviews or user-testing/performance tests. In one instance, presence of the app in the top 100 of a category of the app store was also suggested as a potential proxy for usability.

Appendix 4 Description of quality criteria used in other studies

Quality indicator		Summary definition created for thesis	Examples of app assessment studies where indicator is captured without using a developed tool	Example studies of app quality assessment tools that use this indicator (NB: not all assessing PA apps)
1	Acceptability or Participant perceptions	Positive and negative feedback or recommendations from users on the app content e.g. preferences or recommendations for information, tone or features. Can include user ratings or reviews of the app (the latter may overlap with other indicators such as 10, or 4, these should be coded separately). May include barriers and facilitators (feasibility) to use of the app and/or smartphone such as it being easy to fit self-report requests into your routine, or forgetting to carry the phone. May also include whether or not	User ratings from app stores (Bondaronek et al., 2018) Coders assessed app to see if it was generally recommendable to a client for use to improve health or prevent disease using the question: *"As a health care professional, I would use this app for my personal use or recommend it for use by one of my clients"* (West et al., 2012) Interviews asking for reactions to carrying and using a mobile phone device, the obtrusiveness of the monitoring, reasons for missing survey prompts. (Dunton et al., 2011) Coded user reviews for praise, criticism and recommendations	Assessed user reviews for length, relationship to star ratings and app category. Suggested that they indicate user experience but also ease of use and usefulness, therefore overlapping with other criteria proposed for this thesis (Vasa et al., 2012) Used PsyberGuide evaluation criteria which includes number of consumer ratings (Powell et al., 2016) Suggested using *""lean" principles that specify methods for early-stage testing of features related to feasibility including: acceptability and usability (will the target*

		users/providers/practitioners would recommend the app to others/patients	Focus group to identify preferences for future content. (Milward et al., 2016)	*audience [e.g., patients, Health Care Practitioners] incorporate and sustain the intervention into their lives/clinical practice?)"* and *"practicability (Can it be delivered with minimal burden?)"* (Murray et al., 2016) System Usability Scale asks for rating of *"I think that I would like to use this system frequently"* (Brooke, 1996) Consumer reviews and ratings established as a method for quality assessment of apps (BinDhim et al., 2014) Mobile App Rating Scale captures average and number of user reviews (Stoyanov et al., 2015)
2	Aesthetics	Visual attractiveness of the app interface design in terms of colours, fonts, and layout. How professional the design is. How pleasing to the eye the design and	Coded user reviews for aesthetics. Focus groups to identify preferences for appearance of apps (Milward et al., 2016). Interface design, assessed *"using*	Used the Health-Related Website Evaluation Form which includes an assessment of design including style, fonts, graphics, layering (Pealer and Dorman, 1997; Taki et al.,

		layout is. Can include relevance of design to the behaviour.	*an adapted version of heuristics for user interface design that concentrates on logical and thematic use of colours and fonts, and ease of reading and legibility"* (Reynoldson et al., 2014).	2015) Developed quality assessment tool for apps, which included assessment of interface design. This included questions about design coherence in terms of colour and configuration, icon arrangement and harmony with app, coherent icon categorization. (Jin and Kim, 2015) Mobile App Rating Scale captures app Aesthetics using three questions on layout, graphics and visual appeal (Stoyanov et al., 2015) Quality of Experience of mHealth applications tool assesses appearance using two questions relating to the adequacy of the appearance and whether respondent would change anything (Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013)

3	Behaviour measurement tools (e.g. PA measurement)	The type of tool/method used by the app to measure physical activity (or other behaviours as applicable) and its related validity and reliability. For example, objective or subjective measures. Also can include the definitive accuracy of the PA tools for capturing physical activity or perceived accuracy of in-app calculations.	Assessed *"technologies used to objectively measure physical activity"* including measurement accuracy (Bort-Roig et al., 2014) Captured how physical activity was measured within an app and recommended that reliability and validity are assessed in future (Knight et al., 2015) Assess whether object physical activity assessment method was used (Muntaner et al., 2015)	Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for use of a validated assessment tool to measure behaviour and use of objective measures – this theme was related to evaluations of apps and so it was unclear if it was definitively about measures that the app used, or measures that were used to determine effectiveness of the app however (see theme 7) (McMillan et al., 2015) Quality of Experience of mHealth applications tool assesses whether respondent thinks data are reliable (under Content Quality) and if the calculations done by the application are correct (Under Precision) (Martínez-Pérez et al., 2013)
4	Credible,	Content of the app is likely	Adapted the WHO quality outcome	Developed Currency

	trustworthy/appropriate or useful/essential information	to be accurate or believable – not making impossible or implausible claims. Content is safe for users, won't harm them or will minimise harm or provides a caveat for medical information that requires seeing a professional. Information/app appears useful.	dimensions to assess apps for safety: *"delivering health care which minimizes risks and harm to service users"* (Peiris et al., 2014) Coders assessed if app was credible or trustworthy (West et al., 2012) Clinical content and use of a particular pain assessment mnemonic routine used by clinicians were assessed (Reynoldson et al., 2014)	Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on relevance: *"Is the level of information appropriate for the app's target audience? Does the app do what you'd expect based on its description?"* (McNiel and McArthur, 2016) Developed quality assessment tool via review of existing apps for medication-related problems. Included assessment of whether app included medical disclaimer that it doesn't replace a health care professionals judgement (see Section B) (Loy et al., 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for whether or not they "*aimed at the right level for the target population"* and if they were

				"*likely to cause harm"* to users and if the app is "*"Information Standards" certified"* (see themes 3, 4, 7) (McMillan et al., 2015) Developed quality assessment tool for apps, which included assessment of accuracy of content including whether information provided is accurate; and objectivity including whether professional healthcare information is provided, provided systematically, information is from cited, authoritative sources and/or medical experts (Jin and Kim, 2015) Predefined list of what app should contain (essential content) was established as a method for quality assessment of apps (BinDhim et al., 2014) Quality of Experience of mHealth applications tool assesses usefulness of app

				with one question: "*can you do the same without the application?*" (under Content Quality) (Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013) Mobile App Rating Scale captures app Information consisting of seven questions on accuracy of description, achievable goals, quality and quantity of information, visual information, legitimacy of source and evidence base (Stoyanov et al., 2015)
5	Currency/Maintenance of the app and its documentation	The date of the last update/regularly updated (no consensus in the literature, range between 1 and 6 months since last update) and date of creation and last update are reported. App documentation is updated as well as the app itself.	Date of last update (Reynoldson et al., 2014) Last update in app store (Bondaronek et al., 2018)	An adapted version of the Silberg Scale was used to assess information quality in apps including Currency; whether the app was modified in the previous month and what the creation and last-modification data were (Griffiths and Christensen, 2000; Jeon et al., 2014; Silberg, 2011) Developed Currency

				Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on currency: *“When was the app developed? How recently was the app last updated?”* (McNiel and McArthur, 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps to see if *“Documentation updated along with app updates?”* (see theme 8) (McMillan et al., 2015) Used the Health-Related Website Evaluation Form to assess apps for data last updated (Pealer and Dorman, 1997; Taki et al., 2015) Mobile App Rating Scale captures last update (Stoyanov et al., 2015) Quality of Experience of mHealth applications tool

				assesses currency with one question: *“Does the application receive updates regularly?”* (under Content Quality) (Martínez-Pérez et al., 2013)
6	Development process and teams	Affiliations or credentials of app development team (University, Industry, Government, Commercial or Non-commercial etc.), involvement of experts and users in development.	Developer, whether or not they were a team or an individual, professional background, whether or not they worked with relevant experts and whether or not they'd created any other apps (Reynoldson et al., 2014) Expert involvement, user involvement and organisational affiliation. (Bondaronek et al., 2018)	An adapted version of the Silberg Scale was used to assess information quality in apps including Authorship and Disclosure; author names, affiliations and credentials, application ownership and sponsorship (Griffiths and Christensen, 2000; Jeon et al., 2014; Silberg, 2011) Developed Currency Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on authority: *“Who is the app's developer/creator/sponsor? What are the developer's credentials or organizational affiliations?”* (McNiel and McArthur, 2016)

				Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for if they were "*Developed in collaboration with target group?"* and if health professionals were involved in development and if the app was piloted (see theme 2) (McMillan et al., 2015) Used the Health-Related Website Evaluation Form to assess apps for authorship, profession, experience, qualification and education of author (Pealer and Dorman, 1997; Taki et al., 2015) *"Availability of authors names, credentials and level of healthcare professional involvement in the app development"* established as a method for quality assessment of apps (BinDhim et al., 2014) Used PsyberGuide evaluation

				criteria which includes source of funding for research (Powell et al., 2016)
7	Effectiveness/Potential impact	Evidence of improvements in relevant health outcomes/behaviours. Potential for impact on behaviour or health (can be measured by tools such as the Precede-Proceed Model which includes judgements of whether or not the intervention addresses predisposing, enabling or reinforcing factors that are thought to determine behaviour) (Has been interpreted as use of evidence-based content or theory (Bondaronek et al., 2018), but that is captured under items 9 and 12)	Advocated efficacy testing of apps through various methods such as RCTs (Jake-Schoffman et al., 2017) Coders assessed app *“according to its level of anticipated influence to potentially change behaviour”* informed by the Precede-Proceed Model and predisposing, enabling and reinforcing factors evidenced in the app, namely app content (West et al., 2012) Adapted the WHO quality outcome dimensions to assess apps for effectiveness: *“delivering health care that is adherent to an evidence base and results in improved health outcomes for individuals and communities, based on need”* (Peiris et al., 2014)	Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for any evaluation performed/proposed including evaluation of efficacy, if the evaluation was done by an independent body, with specialist input and when the evaluation was planned for (see theme 7) (McMillan et al., 2015) Used Anxiety and Depression Association of America evaluation criteria which includes rating apps on perceived effectiveness and research evidence (Powell et al., 2016) Mobile App Rating Scale captures potential impact of app via 6 questions on the likelihood of the app to raise

				awareness, increase knowledge, change attitudes, increase intention to change, prompt help-seeking and change behaviour (Stoyanov et al., 2015)
8	Engagement	Use of methods to encourage user interactivity with the app, can include use of certain strategies or features that promote/inhibit for example, feedback, tailoring, prompts/reminders, gamification. (Often this term has also been used to refer to usage/response to app intervention features or feasibility, such as required step-count submissions or required message responses (Monroe et al., 2015). Or it has encompassed both these and the above summary definition (Rose et al., 2017). These items have	Assessed participant engagement in terms of features/content that facilitated or hindered using the smartphone (Bort-Roig et al., 2014)	Used an adapted version of health behaviour theory-based website assessment instrument (Doshi et al., 2003). Included assessment of five levels of user interaction, from provision of information and guidelines to user, to assessment of current behaviour, to feedback on behaviour, to non-individualised assistance, to individually tailored assistance (West et al., 2013; Cowan et al., 2013) Created a chronic disease app evaluation checklist based on other checklists and qualitative studies including Engagement comprised of questions on gamification, customisation, interactivity, engagement

		been separated out into other criterion here).		through use of plug-ins, encouraging self-awareness and behaviour change (Anderson et al., 2016) Developed Currency Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on other factors: *“Is the app fun or interesting to use?”* (McNiel and McArthur, 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for if they “*Tailored intervention based on [user] responses”* and *“user progress”* (see theme 4) (McMillan et al., 2015) Mobile App Rating Scale captures engagement using 5 questions on whether the app is entertaining, interesting, allows customisation, is interactive, and if it’s

				appropriate for the target group (Stoyanov et al., 2015) Used Anxiety and Depression Association of America evaluation criteria which includes rating apps on personalisation and interactiveness /feedback (Powell et al., 2016)
9	Evidence-based content/components	Use of techniques, strategies, information, practice or recommendations that are based on scientific evidence that demonstrates their usefulness. This can include behaviour change techniques or ‘predictors’ associated with improvements in behaviours in general or the target behaviour, prescribed behaviours or practices advocated by reliable government bodies such as Public Health England, or the National Institute of Health. Techniques can be	*“Presence or absence of techniques used to promote change were assessed,”* derived from reviews (Bardus et al., 2016) Apps coded for *“adherence to 13 evidence-informed practices for weight control”* (Breton et al., 2011) Presence of Behaviour Change Techniques in apps (Direito et al., 2014) Coded presence or absence of recommendations for physical activity - if they included *“150 minutes/weekly of moderate to vigorous intensity physical activity or ≥2 days/weekly of whole body*	Developed Currency Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on accuracy: *“Where does the information used in the app come from? Is the information supported by evidence?”* (McNiel and McArthur, 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for if behaviour change techniques were employed (see theme 5) (McMillan et al., 2015)

		captured by referring to taxonomies such as the Behaviour Change Taxonomy v1 (Michie et al., 2013)	*strengthening activities"* (Knight et al., 2015) Advocated evaluating evidence-based content via comparisons to clinical guidelines, use of evidence-based protocols or use of behaviour change techniques (Jake-Schoffman et al., 2017) Presence of "*behavioural strategies included in evidence-based weight loss interventions",* used in apps (Pagoto et al., 2013)	The inclusion of evidence-based content established as a method for quality assessment of apps (BinDhim et al., 2014)
10	Functionality/Usability	Ease of use of the app and/or smartphone features, such as navigation, terminology, design in relation to ease of use, not aesthetics (see 2) as well as general perception of how much support might be required for use or how complex or inconsistent it might be. Functional errors related to app operations such as bugs/crashing also captured here. Includes practicality of	Reported studies of mobile technologies that evaluated usability through interviews, questionnaires, tasks and reporting 'ease of use' (O'Reilly and Spruijt-Metz, 2013) Reported qualitative studies of apps reporting usability as an issue, collected via interviews and questionnaire (Coughlin et al., 2016) Coded user reviews for functionality. Focus groups to	Created a chronic disease app evaluation checklist based on other checklists and qualitative studies including Functionality which included questions on use of health warnings when readings are out of range, tactile, visual or sound-based 'feedback' [alerts] structural navigation, intuitive design, connection to services and performance power (Anderson et al., 2016) Developed Currency

		use for promoting or capturing physical activity based on functions and features. Can be assessed by questionnaires such as the System Usability Scale, interviews or user-testing/performance tests. In one instance, presence of the app in the top 100 of a category of the app store was also suggested as a potential proxy for usability.	determine preferences for functionality of apps. (Milward et al., 2016) Ease of use assessed including navigation, content, accessibility, interactivity and connectivity derived from heuristics for evaluating mobile commuting usability flaws (Reynoldson et al., 2014) Reported on studies that found apps/features easy to use (but defined this under acceptability) (Monroe et al., 2015) Advocated usability testing via lab-based testing, field testing, user feedback from reviews and ratings (Jake-Schoffman et al., 2017)	Relevance Authority Accuracy Purpose (CRAAP) assessment tool for apps. Included questions on other factors: *"Is the app easy to use?"* (McNiel and McArthur, 2016) Developed quality assessment tool via review of existing apps for medication-related problems. Included assessment of usability based on whether app is intuitive to use, any interface problems, organised layout with readable font and understandable terms, includes a help feature, if it has any broken links and frequency of crashing (see Section C) (Loy et al., 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for *"usability"* including *"special features for those with specific needs"* (see theme 3) (McMillan et al., 2015)

				Developed quality assessment tool for apps, which included assessment of understandability, suitability of design and accuracy of wording including questions on whether app information is readily understandable, explained in everyday terms, easy to read in general, has organised and logical content, icon meanings are clear, readable clear app structure, concise and precise instructions and accurate spelling and grammar (Jin and Kim, 2015) Used Anxiety and Depression Association of America evaluation criteria which includes rating apps on ease of use (Powell et al., 2016) Quality of Experience of mHealth applications tool assesses Ease of use using 2 questions relevant to the thesis

				definition: *“Do you find what you need? Do you think that the traditional method used so far is more difficult or does not exist?”* It also assesses Performance using 2 questions *“Do you think you might have a more optimized performance? Do you find some kind of error or problem while using the application?”* Finally, it assesses Learning*: “Do you think that the time for learning the use of the application is appropriate?”* (Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013) Usability of predefined app functions established as a method for quality assessment of apps (BinDhim et al., 2014) System Usability Scale asks for rating of *“I thought the system was easy to use”* and *“I found the system unnecessarily complex”* and *“I think I would need the support*

				of a technical person to be able to use this system" and *"I found the various functions in the system were well integrated"* and *"I thought there was too much inconsistency in this systems"* and *"I would imagine that most people would learn to use this system very quickly"* and *"I found the system very cumbersome to use"* and *"I felt very confident using the system"* and *"I needed to learn a lot before I could get going with this system"* (Brooke, 1996) Mobile App Rating Scale captures Functionality using 4 questions about app performance, ease of use, navigation and gestural design (Stoyanov et al., 2015)
11	Security/Privacy	Data privacy and/or security. For example, could include availability and accessibility of a privacy policy as well as its content, or a required login for the	Based on recommendations of Information Commissioners Office, 8 items asking if app has a privacy policy, if its available without app download, available after app download, if it has a short form	Created a chronic disease app evaluation checklist based on other checklists and qualitative studies including Information Management comprised of questions on privacy and data

		app or ability to make personal content private rather than sharing with app community. Could include meeting Data Protection standards.	notice of the policy, if the policy is in other languages, if the app collects personally identifiable information and if data is shared with a 3rd party (Bondaronek et al., 2018) Focus group to determine preferences for apps included theme on autonomy and privacy of online communities in apps (Milward et al., 2016) Security assessed as part of ease of use – queries if data is kept private/safe for example via encryption, and if it can be backed up and restored (Reynoldson et al., 2014)	security (Anderson et al., 2016) Developed quality assessment tool via review of existing apps for medication-related problems. Included assessment of how app states Privacy Policy, if users can choose content to share and with whom (see Section D) (Loy et al., 2016) Developed tool based on adherence to NICE behaviour change guidance. Included assessment of apps for if they "*comply with Data Protection standards"* (see theme 9) (McMillan et al., 2015) Developed quality assessment tool for apps, which included assessment of security including information about a Privacy Policy, security policies and explains its security system (Jin and Kim, 2015)

				Quality of Experience of mHealth applications tool assesses security using 2 questions: *“Do you think that this application has appropriate security methods to protect data that are introduced?”* and *“Do you think that the data obtained with this application are sufficiently protected?”* (Martinez-Perez et al., 2015; Martínez-Pérez et al., 2013)
12	Theoretical underpinning/components	Use of theoretical constructs from established behaviour change theories, or mention of use of theory, to inform development/content/ evaluation of app in some way. Has/can be assessed by standardised tools such as the one developed by Doshi et al., (2003).	Assessed for whether or not it was based on theoretical guidelines (Muntaner et al., 2015)	Used an adapted version of health behaviour theory-based website assessment instrument (Doshi et al., 2003). Included assessment of constructs from four major theories of behaviour change across 5 levels of user-interaction (Cowan et al., 2013; West et al., 2013) Assessed using Theory Coding Scheme from Michie and Prestwich, (2010) (Roberts et al., 2017)
13	Usage/Compliance	Not to be confused with engagement, usage or	Participant usage, referred to as compliance, frequency of use over	Developed tool based on adherence to NICE behaviour

		compliance refers to responses to app content such a required step-count submissions or response to prompts to complete questions. Similar to fidelity.	time and what features or supplements influenced usage (Bort-Roig et al., 2014). Reported studies that assessed 'engagement' but is reporting usage as defined by this thesis: amounts of required data entry completed, responses to study messages (Monroe et al., 2015) Reported studies that assessed 'engagement' but is reporting usage at times, as defined by this thesis e.g. number of times logged in per week (Rose et al., 2017)	change guidance. Included assessment of apps for any evaluation performed/proposed including uptake and reach and fidelity (see theme 7) (McMillan et al., 2015) Suggesting using *""lean" principles that specify methods for early-stage testing of features related to feasibility including: demand (Will relevant stakeholders use it?)"* (Murray et al., 2016)

Appendix 5 PROSPERO protocol for systematic review

PROSPERO
International prospective register of systematic reviews

NHS National Institute for Health Research

Quality and characteristics of physical activity apps that include feedback on affect/mood/emotion: a systematic review of the literature

Laura Lamming, Andy Scally, Neil Small, Maria Horne, Ian Kellar

Citation

Laura Lamming, Andy Scally, Neil Small, Maria Horne, Ian Kellar. Quality and characteristics of physical activity apps that include feedback on affect/mood/emotion: a systematic review of the literature. PROSPERO 2018 CRD42018107289 Available from:
http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018107289

Review question

1) Are there any physical activity promotion apps in the academic literature that include information on user affect/mood/emotion as a component?

2) What are the characteristics of apps in general and specifically, the information on affect/mood/emotion?

3) What evidence is there of their quality?

Searches

- CINAHL

- EMBASE

- MEDLINE (On Ovid)

- PsycINFO

- Scopus

The searches will be limited to publications in the English language and publications from the year 2000 onwards. Where applicable, type of publication will be limited to academic journals, theses/dissertations, articles in press, conference papers, but not abstracts.

Types of study to be included

It is anticipated that experimental, observational, (e.g. cohort studies), exploratory or correlational, qualitative (where an app is being developed and/or assessed for acceptability, feasibility, usability etc.) and mixed-methods studies will be eligible. Evaluations of publicly available apps will also be included. Reviews will be excluded, but retained in order to search their reference lists for other relevant app studies. Protocols of relevant interventions (apps) will be retained in order to identify other potentially relevant studies in progress, at which point, relevant papers will be located or authors will be contacted for details.

Condition or domain being studied

Physical activity promotion in free-living adults.

Participants/population

Free-living adults (18 y.o+). Apps must not be targeted at a specific clinical population.

Intervention(s), exposure(s)

Intervention studies (where app is used), development of interventions (apps), or evaluation of commercial apps exploring uptake, maintenance or promotion of, (lifestyle or structured) physical activity behaviour or decreasing sedentary behaviour/amount of inactivity are potentially eligible. Apps must also include, but aren't necessarily limited to, providing information about participant affect/mood/emotion (positive or negative), preferably with a focus on physical activity-contingent affect.

Comparator(s)/control

Any control group will be eligible for inclusion, including having no control group.

Context

The main outcome is to examine the (as yet unrealised) potential for a preventative intervention to increase physical activity levels. Studies investigating a physical activity promotion app that includes information on affect/mood/emotion in free-living adult populations are of interest. We hypothesise that such interventions/apps could be optimised in terms of evidence-based content and quality and administered on a wide scale, at little cost and with benefits for both uptake and maintenance of physical activity.

Main outcome(s)

Characteristics of physical activity promotion apps which provide information on affect/mood/emotion

Characteristics of the information or feedback provided on affect/mood/emotion

Quality of app captured by multiple outcomes including but not limited to: Mobile App Rating Scale, Theoretical Coding Scheme, behaviour change techniques and qualitative outcomes

These outcomes will demonstrate the quality of existing physical activity apps that attempt to use short term-benefits (mood outcomes) as motivation, both in terms of evidence-based content and usability which can be used to facilitate choice of apps by potential users and/or clinicians who wish to recommend physical activity apps to patients. Such apps may be advisable for those who have previously struggled to engage with interventions/activity promotion that focuses on long-term outcomes.

Additional outcome(s)

None.

Data extraction (selection and coding)

One researcher will initially screen titles for inclusion. Each abstract and full text will be screened independently by two researchers. Any discrepancies will be resolved by discussion with an author.

Using a standardised piloted proforma, the following data will be extracted independently by two authors and discrepancies resolved by a third author:

• Administrative details of paper (author, title, year, country study was conducted in)

• Methods (study type, design, duration)

• Participants (number, setting, diagnostic criteria, age, sex, country, socio-demographics, ethnicity, specifically, details of target group for app)

• Interventions (number of groups, specific intervention - with focus on details of app – name, description, type of PA targeted, platform availability, developers if not authors, theoretical underpinnings, presence of PA recommendations for adults, and if in an app store: cost, size, store ranking, number of ratings and version, presence of behaviour change techniques, fidelity, presence, stage, details of involvement and demographic details of any users involved in development/testing. Wherever possible the app will be downloaded for review over a 48 hour period to further assess content and details).

• Outcomes (Detailed overview of quality outcomes such as usage, usability and acceptability outcomes including any other outcomes that seem relevant to use, uptake and maintenance. Any other measures/tools used to determine app quality and definitions of units if applicable.)

• Results (Quality outcomes such as usage/usability/acceptability or other quality related data: all available information present in paper. Any other app quality as measured by tools or otherwise and whether or not quality issues were addressed.)

• Other (miscellaneous comments and key conclusions of relevance to the review questions, references to other relevant studies or linked papers that describe reported apps further.)

Risk of bias (quality) assessment
As the primary focus of the review is on identifying apps and assessing their quality a formal assessment of study quality related to design and reporting will not be performed.Instead, quality of the apps themselves will be assessed by multiple outcomes including but not limited to: Mobile App Rating Scale, Theoretical Coding Scheme, behaviour change techniques and qualitative outcomes. Quality assessment of apps will be assessed by two researchers and descrepancies resolved by a third where necessary. Quality of the apps will not influence the planned synthesis.

Quality and Risk of Bias Checklist for Studies That Review Smartphone Applications will be used to ensure reporting of methods of the app review (BinDhim et al., 2014).

Strategy for data synthesis
Interrater reliability will be calculated for MARS scores where they are generated, and BCT coding, percentage agreements and ordinal and nominal Krippendorffs alpha (kalpha) will be used for both in accordance with previous app evaluations. Krippendorffs alpha is considered a more versatile alternative to Cohens kappa and calculates disagreement rather than agreement, which, among other features, is thought to make it more reliable than other coefficients.

Findings will be tabulated and a narrative synthesis (Popay.et al 2006) will be conducted. Depending on findings, results may be divided by publicly available (in app stores) and restricted apps (available via researcher or other specialist only).

Analysis of subgroups or subsets
If sufficient data are available, separate narrative analysis will be performed for individuals with different socioeconomic status (groups to be determined) and for apps that are, or are not, publicly available in an app store.

Contact details for further information
Laura Lamming
L.Lamming1@bradford.ac.uk

Organisational affiliation of the review
University of Bradford

Review team members and their organisational affiliations
Miss Laura Lamming. University of Bradford
Mr Andy Scally. University of Bradford
Professor Neil Small. University of Bradford
Assistant/Associate Professor Maria Horne. University of Leeds
Assistant/Associate Professor Ian Kellar. University of Leeds

Type and method of review
Narrative synthesis, Systematic review

Anticipated or actual start date
01 February 2018

Anticipated completion date
30 April 2019

Funding sources/sponsors
None.

Conflicts of interest

Language
English

Country
England

Stage of review
Review Ongoing

Subject index terms status
Subject indexing assigned by CRD

Subject index terms
Affect; Emotions; Exercise; Feedback; Humans

Date of registration in PROSPERO
31 October 2018

Date of publication of this version
31 October 2018

Details of any existing review of the same topic by the same authors

Stage of review at time of this submission

Stage	Started	Completed
Preliminary searches	Yes	Yes
Piloting of the study selection process	Yes	No
Formal screening of search results against eligibility criteria	No	No
Data extraction	No	No
Risk of bias (quality) assessment	No	No
Data analysis	No	No

Versions
31 October 2018

Appendix 6 Systematic review search strategies

CINAHL (EBSCO)

1. AB walk* OR TI walk*
2. AB exercise* OR TI exercise*
3. (MH “Exercise+”)
4. (AB “physical activity”) OR (TI “physical activity”)
5. (MH “Physical activity”)
6. (AB “physical* fit*”) OR (TI “physical* fit*)
7. (MH “Physical Fitness+”)
8. (AB inactivit* OR inactive) OR (TI inactivit* OR inactive)
9. **1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8**
10. AB mood OR TI mood
11. (AB affect or affective*) OR (TI affect OR affective*)
12. (AB feeling OR feelings) OR (TI feeling OR feelings)
13. AB “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
14. TI “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
15. (AB emotion OR emotions OR emotional) OR (TI emotion OR emotions OR emotional)
16. (AB self-efficacy OR “self efficacy”) OR (TI self-efficacy OR “self efficacy”)
17. AB circumplex OR TI circumplex
18. **10 OR 11 OR 12 OR 13 OR 15 OR 16 OR 17**
19. AB smartphone* OR TI smartphone*
20. (MH “Smartphone”)
21. (MH “Mobile applications”)
22. (AB “mobile phone*”) OR (TI “mobile phone*)
23. (AB apps OR app OR application*) OR (TI apps OR app OR application*)
24. (AB Iphone OR I-phone OR android OR iOS) OR (TI Iphone OR I-phone OR android OR iOS)
25. (AB “mobile health”) OR (TI “mobile health”)
26. (AB phone OR mobile OR telephone) OR (TI phone OR mobile OR telephone)
27. (AB “Mobile device*”) OR (TI “mobile device*”)
28. AB “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”
29. TI “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”

30. (AB “cell phone*” OR “cellular phone”) OR (TI “cell phone*” OR “cellular phone”)
31. (MH “Cellular Phone+”)
32. AB digital OR TI digital
33. **19 OR 20 OR 21 OR 22 OR 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31 OR 32**
34. **9 AND 18 AND 33 (English)**
35. **Limiters – Date of Publication 20000101-20181231.**

PsycINFO (EBSCO)

1. AB walk* OR TI walk*
2. DE “Walking”
3. AB exercise* OR TI exercise*
4. DE “Exercise” OR DE “Aerobic Exercise” OR DE “Weightlifting” OR DE “Yoga”
5. (AB “physical activity”) OR (TI “physical activity”)
6. DE “Physical activity”
7. (AB “physical* fit*”) OR (TI “physical* fit*)
8. DE “Physical Fitness”
9. (AB inactivit* OR inactive) OR (TI inactivit* OR inactive)
10. **1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9**
11. AB mood OR TI mood
12. (AB affect or affective*) OR (TI affect OR affective*)
13. (AB feeling OR feelings) OR (TI feeling OR feelings)
14. AB “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
15. TI “feel* N0 (state*OR states OR good OR better OR positive OR inventory OR change* OR scale*)
16. (AB emotion OR emotions OR emotional) OR (TI emotion OR emotions OR emotional)
17. (AB self-efficacy OR “self efficacy”) OR (TI self-efficacy OR “self efficacy”)
18. AB circumplex OR TI circumplex
19. **11 OR 12 OR 13 OR 15 OR 16 OR 17 OR 18**
20. AB smartphone* OR TI smartphone*
21. (AB “mobile phone*”) OR (TI “mobile phone*)
22. (AB apps OR app OR application*) OR (TI apps OR app OR application*)
23. (AB Iphone OR I-phone OR android OR iOS) OR (TI Iphone OR I-phone OR android OR iOS)
24. (AB “mobile health”) OR (TI “mobile health”)

25. (AB phone OR mobile OR telephone) OR (TI phone OR mobile OR telephone)
26. (AB “Mobile device*”) OR (TI “mobile device*”)
27. AB “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”
28. TI “ecological momentary assessment” OR EMA OR “ecological-momentary-assessment”
29. (AB “cell phone*” OR “cellular phone”) OR (TI “cell phone*” OR “cellular phone”)
30. DE “Cellular Phones”
31. AB digital OR TI digital
32. **20 OR 21 OR 22 OR 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31**
33. **10 AND 19 AND 32 (English)**
34. **Limiters – Date of Publication 2000-2018**

EMBASE

1. ‘walk*’: ab,ti OR ‘exercis*’:ab,ti
2. ‘physical* fit*’:ab,ti
3. ‘physical activity’: ab, ti
4. ‘inactivit*’:ab, ti OR ‘inactive’:ab,ti
5. ‘exercise’/exp
6. ‘fitness’/exp
7. ‘physical activity’/exp
8. #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7
9. 'mood':ab,ti OR 'affect':ab,ti OR 'affective*':ab,ti OR 'emotion':ab,ti OR 'emotions':ab,ti OR 'emotional':ab,ti
10. (feel* NEAR/1 (state OR states OR good OR better OR positive OR inventory OR change* OR scale*)):ab,ti
11. feeling OR feelings:ab,ti
12. 'self-efficacy':ab,ti OR 'self efficacy':ab,ti OR 'circumplex':ab,ti
13. #9 OR #10 OR #11 OR #12
14. smartphone*':ab,ti OR 'mobile phone*':ab,ti OR 'apps':ab,ti OR 'app':ab,ti OR 'application*':ab,ti OR 'iphone':ab,ti OR 'i-phone':ab,ti OR 'android':ab,ti OR 'ios':ab,ti OR 'mobile health':ab,ti OR 'digital':ab,ti OR 'cell phone*':ab,ti OR 'cellular phone*':ab,ti
15. 'phone':ab,ti OR 'mobile':ab,ti OR 'telephone':ab,ti
16. ('mobile device*':ab,ti OR 'ema':ab,ti OR 'ecological momentary assessment':ab,ti) AND 'ecological-momentary-assessment':ab,ti
17. 'smartphone'/exp
18. 'mobile application'/exp
19. 'mobile phone'/exp

20. **#14 OR #15 OR #16 OR #17 OR #18 OR #19**
21. **#8 AND #19 AND #20 AND ([article]/lim OR [article in press]/lim OR [conference paper]/lim) AND [english]/lim AND [embase]/lim AND [2000-2018]/py**

SCOPUS

((TITLE-ABS-KEY (walk*)) OR (TITLE-ABS-KEY (exercis*)) OR (TITLE-ABS-KEY ("physical* fit*")) OR (TITLE-ABS-KEY ("physical activity")) OR (TITLE-ABS-KEY (inactivit*)) OR (TITLE-ABS-KEY (inactive))) AND ((TITLE-ABS-KEY (mood)) OR (TITLE-ABS-KEY (affect OR affective*)) OR (TITLE-ABS-KEY (emotion OR emotions OR emotional)) OR (TITLE-ABS-KEY (feel* W/0 (state OR states OR good OR better OR positive OR inventory OR change* OR scale*))) OR (TITLE-ABS-KEY (feeling OR feelings)) OR (TITLE-ABS-KEY (self-efficacy OR "self efficacy")) OR (TITLE-ABS-KEY (circumplex))) AND ((TITLE-ABS-KEY ("ecological momentary assessment" OR ema OR "ecological-momentary-assessment")) OR (TITLE-ABS-KEY ("cell phone*" OR "cellular phone*")) OR (TITLE-ABS-KEY (digital)) OR ((TITLE-ABS-KEY (smartphone*)) OR (TITLE-ABS-KEY ("mobile phone*")) OR (TITLE-ABS-KEY (apps OR app OR application*)) OR (TITLE-ABS-KEY (iphone OR i-phone OR android OR ios)) OR (TITLE-ABS-KEY ("mobile health")) OR (TITLE-ABS-KEY (phone OR mobile OR telephone)) OR (TITLE-ABS-KEY ("mobile device*"))))) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ip")) AND (LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010) OR LIMIT-TO (PUBYEAR , 2009) OR LIMIT-TO (PUBYEAR , 2008) OR LIMIT-TO (PUBYEAR , 2007) OR LIMIT-TO (PUBYEAR , 2006) OR LIMIT-TO (PUBYEAR , 2005) OR LIMIT-TO (PUBYEAR , 2004) OR LIMIT-TO (PUBYEAR , 2003) OR LIMIT-TO (PUBYEAR , 2002) OR LIMIT-TO (PUBYEAR , 2001) OR LIMIT-TO (PUBYEAR , 2000))

Appendix 7 Title screening guidelines for systematic review

PICOS and guidelines for title screening

Systematic review of apps for physical activity promotion

Responses

Responses in the spreadsheet, including final eligibility decision, will turn red, green, grey or blue depending on your answers. The colours should help indicate whether the paper should be included, reviewed or excluded – see item 6 below for details.

Colour	Meaning
Red	Criteria has definitely not been met and therefore paper will be excluded
Green	Criteria has been met
Grey	Unclear/insufficient information to be sure that criteria has been met – needs abstract review to confirm
Blue	Paper is either a literature review or a protocol and will need abstract reviewing separately for relevance

A note on reviews/protocols

Please note that reviews or protocols that may look relevant studies/topics are not automatically excluded, but can be retained for further review to see if they may contain or relate to ANOTHER relevant paper that we will source. See criteria 3. Please consider the reviews topic based on the rest of the criteria and if it might contain something relevant.

1. **Obviously a completely irrelevant title/aim?**

Response	Possible reasons
Y (yes)	Nothing to do with changing behaviours at all e.g. *'Whole-body cooling does not compromise muscle oxidative capacity in subjects with multiple sclerosis'* Nothing to do with humans (If yes, can stop screening)
N (no)	Includes reference to behaviour change in some way e.g. '*Text messaging intervention for teens and young adults with diabetes'*

U (unsure)	Can include anything relating to correlates, determinants or predictors of behaviour change at this time. e.g. *'Quantitative study of correlates of physical activity in women from diverse racial/ethnic groups: The Women's Cardiovascular Health Network Project--summary and conclusions'* Anything else you might be unsure about related to this item.

2. **Does it refer to/allude to an app or use any other term that could include an app?**

Response	**Possible reasons**
Y (yes)	Uses the terms: app, digital, technology, quantified self, smartphone application, Ecological Momentary Assessment, Digital diaries, e-Diaries, etc Uses terms that suggest a phone was involved somehow that wasn't restricted to phone call e.g. smartphone, telephone-based, phone, mobile phone etc. These are retained until it can be clarified that an app wasn't involved.
N (no)	Clearly not going to involve an app as a potential intervention tool. e.g. *'Built environment and walking to school: Findings from a student travel behavior survey in Massachusetts'* Involves text-messages with no reference to an app: e.g. *'The effectiveness of text messaging programs on adherence to treatment regimens among adults aged 18 to 45 years diagnosed with asthma: A systematic review protocol'*
U (unsure)	Refers to a text messaging app e.g. Whatsapp etc. If the term 'web' or 'internet' is used or if its not clear whether or not an app may have been involved. Retain if it involves an intervention technique that could at a push be delivered digitally e.g. Motivational Interviewing (It's a stretch, but just in case) Anything else you might be unsure about related to this item.

3. **Is it a review or protocol?**

Response	Possible reasons
Y (yes)	Says review or protocol in the title
N (no)	Doesn’t say review or protocol in the title
U (unsure)	Suspect it might be a review or protocol judging by the way its written

4. **Does it include exploring physical activity in anyway and/or its uptake, maintenance, promotion OR reducing sedentary behaviour?**

Response	Possible reasons
Y (yes)	Clearly looking at promoting some sort of physical activity (all variations on the term relevant) Looking to change/influence/explore multiple behaviours that INCLUDE physical activity
N (no)	Where physical activity it being used purely as substitute or adjunct for smoking cessation/addiction management as these represent a specialist group Looking at rehabilitation or bringing people back to a normal level of functioning. Looking at physical functioning in general or balance and falls. Any therapeutic physical activity/exercise or intervention. Where other behaviours not including physical activity are examined e.g. diet, learning, smoking, alcohol consumption etc. Where physical activity is not mentioned at all and it seems highly unlikely that its incorporated.
U (unsure)	Lifestyle interventions Weight loss/gain prevention interventions (These may be removed in abstract review as I’m not convinced they are necessarily relevant and don’t actually constitute a specialist group like smokers/addicts) Self-management interventions Wellbeing interventions

	All of the above could include a PA component.

5. **Are the participants/target group for change free-living, condition-free, adults?**

Response	Possible reasons
Y (yes)	18 years or older – assume adults if referred to as 'employees' or 'workers' or 'college/university students' or 'veterans' AND Live in the community AND Suffer from no condition/disease. HOWEVER, for the purposes of inclusivity and potential for healthy control groups the following can be included for now: • obese/overweight individuals • smokers
N (no)	Children/teenagers under the age of 18, e.g. high school students Prisoners, hospitalised patients, those in assisted living. Those with a condition or disease including mental or physical health conditions Those who have recently had a disease e.g. 'post stroke' or 'cancer survivors' or 'recently had gestational diabetes'
U (unsure)	Pregnant women Those who are pre-disease e.g. prediabetic, prehypertensive. Those described as 'at risk' of a condition Youth/young adults (unless age confirmed) Any other instance where it's unclear what participant status is

6. Include/Exclude/Review

Response	Possible reasons
I (include)	All answers were green or grey
E (exclude)	At least one answer was red
R (review)	Criteria 3 (review/protocol) was blue AND the rest of the criteria were green or grey

7. Notes

Make a note if:

- Duplicate
- Want to clarify any responses you've made
- You think it's an inappropriate document – book review, book chapter, editorial, conference abstract, not in English.
- Any other concerns you might have

Appendix 8 Abstract screening guidelines for systematic review

PICOS and guidelines for Abstract screening

Systematic review of apps for physical activity promotion

1. **Does it include increasing or exploring uptake, maintenance or promotion of physical activity?**

Response	Possible reason
Y (yes)	Clearly aims to increase structured/unstructured physical activity OR exercise OR reduce sedentary behaviour. Can include papers that report multiple studies of different behaviours as long as PA is targeted. Can include interventions that report exercise-related outcomes such as exercise self-efficacy, as the goal is to find PA apps, not test efficacy. Can aim to increase strength and performance or speed for now. Can include correlates, determinants, predictors and barriers of physical activity for now as they are 'exploring' physical activity.
N (no)	Where it's clear that the aim of the paper has nothing to do with physical activity in humans whatsoever e.g. if it's about animals or reading ability or social development etc – at this point coding can stop. Looking at multiple behaviours that don't include physical activity. Activity is part of rehabilitation, pre-rehabilitation.
U (unsure)	Where its multiple behaviours but may include PA, but it's not clear e.g. 'lifestyle intervention' 'weight loss' 'wellbeing intervention'

2. **Does it refer to a native app or allude to it as an intervention tool that is being developed/tested in paper?**

Response	Possible reason
Y (yes)	Refers to an app(s) for behaviour change/intervention and/or is developing or testing or assessing it in some way. Does not have to be testing app per se – just mention that one was used/developed etc.

	NB: Be careful as some text based interventions still include an app and some activity monitors may include an app but it may not be mentioned in the abstract
N (no)	Survey of health app usage in a population. Uses the web/internet, rather than a specially designed native app. No technology at all is being used – i.e. face to face delivery. If phone calls are being used, rather than apps e.g. if they refer to ‘telephone counselling’. But be careful as telephone is often used but not clear how. If generic texts are being used, rather than apps (see note in ‘Yes’ above) – unless it’s a texting app e.g. Whatsapp, Facebook messenger. If it’s a correlation/determinates study that just uses a survey to look at associations. If EMA is being used purely for self-monitoring/assessment of correlates rather than for PA promotion.
U (unsure)	Where it’s not clear if a native app is being used or if participants are just using the internet through their phone. Retain if social media apps may be being used instead. Retain if EMA/Ecologically Momentary Assessment is used but it’s not clear how – i.e. just assessment or as part of app- it’s being captured as this can be via smartphone. Retain if activity tracker is used but an app isn’t explicitly mentioned as many have connected apps. Retain if a telephone is used but it’s not clear how. I.e. if it states interviews then exclude, but if it’s just ‘telephone based’ that retain for now.

3. **Are the participants/target group for change/exploration adults, free living and non-clinical?**

Response	**Possible reason**

Y (yes)	Participants are 18 years old+ AND Live in the community AND Are not part of a clinical disease/condition group Adults (with no age specified) For now, include if they might be ‘pre-diabetic’ or ‘pre-hypertensive’ ‘at risk’ and pregnant, overweight or obese. Include if participants are just described as ‘college students’ or ‘university students’ or ‘workers’, ‘employees’, in the first instance. Smokers if the aim is just to get them more active but not stop smoking ‘Older adults’ unless age suggests less than 18 Any particular professions that are referred to that require staff to be adults e.g. ‘nurses’ (with no age specified)
N (no)	Less than 18 years old Live in assisted living, are in hospital, are in prison Have a disease or condition that would therefore warrant specialist investigation. Includes people who have just survived a condition e.g. ‘cancer survivors’ or who recently had a condition. Alcoholics or other addicts Not humans e.g. animals If it involves dyads where the intervention is administered to an eligible population but outcome is measured in an ineligible population e.g. parents and children, where the outcome is measured in children
U (unsure)	Where it’s not clear the status of the participants ‘Women’ or ‘men’ without ages specified ‘Young adults’ or ‘youth’ without age specified

Include/Exclude/Review

Response	Reason
Include	All codes are green
Exclude	At least one code is red OR it's an inappropriate document type (see Notes below)
Review	Codes are either all grey or a mix of grey and green (this will be the case for possible reviews)

Notes

If a document is not appropriate – e.g. it's a book review, duplicate etc., make a note here and exclude.

If it's an erratum or correction to an INCLUDED/INCLUDABLE paper then retain and make a note here.

Can also note if paper is excluded but likely to be relevant to background/discussion for thesis.

Make a note if a paper is reporting analysis of an undescribed intervention that may need to be reviewed for relevance.

Make a note if it's a literature review and classify as Review

Protocols and Reviews

For protocols, use the criteria above.

For reviews, we are trying to find any included studies that may refer to an app so please use the following adaptations to the above criteria:

Question	Revised criteria for review
1. Is it focused on exploring, increasing uptake, maintenance or promotion of lifestyle or structured physical activity behaviour, OR decreasing sedentary behaviour/inactivity ONLY? - Not rehab/prehab! (if completely irrelevant, stop screening)	Multiple behaviours are allowed as PA may be dealt with individually
2. Does it refer to a native app, or allude to it as an intervention that is being tested/developed/assessed in the paper?	Unless it's clear that apps are not being looked at e.g. it reviews face to face counselling, then retain so review can be checked for possibly relevant studies

3. Are the participants/target group for change 18 years+, free-living and non-clinical sample?	Unless it’s clear that its focused on a particular ineligible group only, retain as it may include studies of eligible populations

Appendix 9 Full text screening guidelines for systematic review

PICOS and guidelines for Full text screening

Systematic review of apps for physical activity promotion

1. **Does it include increasing or exploring uptake, maintenance or promotion of physical activity?**

Response	Possible reason
Y (yes)	Clearly aims to increase structured/unstructured physical activity OR exercise OR reduce sedentary behaviour. Can include papers that report multiple studies of different behaviours as long as PA is targeted. Can include interventions that report exercise-related outcomes such as exercise self-efficacy, as the goal is to find PA apps, not test efficacy. Can aim to increase strength and performance or speed for now. Can include correlates, determinants, predictors and barriers of physical activity for now as they are 'exploring' physical activity.
N (no)	Where it's clear that the aim of the paper has nothing to do with physical activity in humans whatsoever e.g. if it's about animals or reading ability or social development etc. – at this point coding can stop. Looking at multiple behaviours that don't include physical activity. Activity is part of rehabilitation, pre-rehabilitation.
U (unsure)	Where its multiple behaviours but may include PA, but it's not clear e.g. 'lifestyle intervention' 'weight loss' 'wellbeing intervention'

2. **Does it refer to a native app or allude to it as an intervention tool that is being developed/tested in paper?**

Response	Possible reason
Y (yes)	Refers to an app(s) for behaviour change/intervention and/or is developing or testing or

	assessing it in some way. Does not have to be testing app per se – just mention that one was used/developed etc. Refers explicitly to it as an app and is definitely not a web application, or pictures and introduction/discussion support the idea that it's a native app e.g. discussion of 'mobile phone intervention' or mHealth or other apps. NB: Be careful as some text based interventions still include an app and some activity monitors may include an app but it may not be mentioned in the abstract
N (no)	Survey of health app usage in a population. Uses the web/internet, rather than a specially designed native app. No technology at all is being used – i.e. face to face delivery. If phone calls are being used, rather than apps e.g. if they refer to 'telephone counselling'. But be careful as telephone is often used but not clear how. If generic texts are being used, rather than apps (see note in 'Yes' above) – unless it's a texting app e.g. Whatsapp, Facebook messenger. If it's a correlation/determinates study that just uses a survey to look at associations. If EMA is being used purely for self-monitoring/assessment of correlates rather than for PA promotion. Just a podcast
U (unsure)	Where it's not clear if a native app is being used or if participants are just using the internet through their phone. Retain if social media apps may be being used instead. Retain if EMA/Ecologically Momentary Assessment is used but it's not clear how – i.e. just assessment or as part of app- it's being captured as this can be via smartphone.

	Retain if activity tracker is used but an app isn't explicitly mentioned as many have connected apps. Retain if a telephone is used but it's not clear how. I.e. if it states interviews then exclude, but if it's just 'telephone based' that retain for now. If it's not clear from description that it definitely doesn't include an app, for example, if it's not clear that the phone was a smartphone or not – make a note to contact authors. If the 'app' is delivered via a PDA or 'PocketPC' Doesn't use the term 'app' or 'application' to refer to it, but images and other text suggests it might be an app. Unsure if it's actually a web application.

3. **Does it involve provision of/feedback on affect/mood/emotion as part of the intervention in any way?**

Response	**Possible reason**
Y (yes)	App provides some sort of information or feedback on participants or users mood/emotion/affect – doesn't necessarily have to be physical-activity contingent affect but it shouldn't be explicitly related to another behaviour e.g. diet NB: Enjoyment in general, or enjoyment of the exercise activity, happiness, pleasant or unpleasant feelings, mental wellbeing, emotional function, non-clinical stress/anxiety could constitute as 'affect/emotion/mood'. NB: Mood that is fed back can have been captured during or immediately after exercise, or any time during the day.
N (no)	No information on participants/users mood/affect/emotion provided by app If mood is measured, but no information about it is provided i.e. it's definitely not available within the app/feedback to the user, once it's been collected. If it's unclear whether or not feedback is provided, code as UNSURE and make a note in the notes section. Clinical mood e.g. clinical depression, or physical wellbeing, is being captured and fed back.
U (unsure)	Insufficient information, but suggests it might be

	worth contacting authors to confirm whether or not feedback on mood was provided. If mood is measured but there does seem to be some feedback or messages received, but there's too little detail to determine if mood feedback was included – make a note to contact authors. If feedback is provided in another tech-based way, not via the app e.g. via a related website If it includes use of a wearable e.g. Fitbit, Jawbone Up and/or mood data is collected/feedback via the wearable instead of the app. If it's not clear if what is being captured should be classified as 'pure' mood/affect/emotion – e.g., affective attitude, mood-based expected outcomes Any other reason to be unsure

4. **Are the participants/target group for change/exploration adults, free living and non-clinical?**

Response	**Possible reason**
Y (yes)	Participants are 18 years old+ AND Live in the community AND Are not part of a clinical disease/condition group Adults (with no age specified) For now, include if they might be 'pre-diabetic' or 'pre-hypertensive' 'at risk' and pregnant, overweight or obese. Include if participants are just described as 'college students' or 'university students' or 'workers', 'employees'. Smokers if the aim is just to get them more active but not stop smoking 'Older adults' unless age suggests less than 18 Any particular professions that are referred to that require staff to be adults e.g. 'nurses' (with no age specified)

	If an average age is given rather an range, that indicates predominantly adult participants (18+)
N (no)	Less than 18 years old Live in assisted living, are in hospital, are in prison Have a disease or condition that would therefore warrant specialist investigation. Includes people who have just survived a condition e.g. ‘cancer survivors’ or who recently had a condition. Alcoholics or other addicts Not humans e.g. animals If it involves dyads where the intervention is administered to an eligible population but outcome is measured in an ineligible population e.g. parents and children, where the outcome is measured in children
U (unsure)	Where it’s not clear the status of the participants but potentially could request information from authors e.g.: ‘Women’ or ‘men’ without ages specified ‘Young adults’ or ‘youth’ without age specified

5. **Is it an appropriate paper?**

Response	**Possible reason**
Y (yes)	Journal publication Thesis
N (no)	Editorial Review Discussion piece Abstract only Book review Duplicate
U (unsure)	Conference papers that are more than just an abstract

Include/Exclude/Review

Response	**Reason**
Include	All codes are green
Exclude	At least one code is red OR it’s an inappropriate document type (see Notes below)

Review	Codes are either all grey or a mix of grey and green (this will be the case for possible reviews)

Notes

If a document is not appropriate – e.g. it's a book review, duplicate etc., make a note here and exclude.

If it's an erratum or correction to an INCLUDED/INCLUDABLE paper then retain and make a note here.

Can also note if paper is excluded but likely to be relevant to background/discussion for thesis.

Make a note if a paper is reporting analysis of an undescribed intervention that may need to be reviewed for relevance – i.e. need to go find another paper.

If it reviews/tests a commercial app

If it's not clear whether or not an app/feedback on mood is occurring and therefore will need to contact the authors. Or if for any other reason, authors should be contacted for clarification.

If there are papers referenced (for example in a review or an ineligible editorial) that should be checked for relevance.

Appendix 10 Search strategies for eligible app studies for systematic review

App and reference	Database and date search conducted	Search strategy
Health Mashups (Bentley et al., 2013)	CINAHL (11th April 2019)	TX “health mashups” OR TX “health mash-ups”
	EMBASE (12th April 2019)	‘health mashups’ OR ‘health mash-ups’
	Medline (12th April 2019)	TX “health mashups” OR TX “health mash-ups”
	PsycINFO (12th April 2019)	TX “health mashups” OR TX “health mash-ups”
	SCOPUS (12th April 2019)	(ALL(“health mashups”) OR ALL (“health mash-ups”))
Motimate (Brindal et al., 2016)	CINAHL (11th April 2019)	TX Motimate
	EMBASE (12th April 2019)	motimate
	Medline (12th April 2019)	TX Motimate
	PsycINFO (12th April 2019)	TX Motimate Limiters – Publication Year: 2000-2018
	SCOPUS (12th April 2019)	ALL (motimate)
ATHENA (Fahim et al., 2014)	CINAHL (11th April 2019)	TX ATHENA AND AU Lee
	EMBASE (12th April 2019)	1. ATHENA 2. Lee:au OR Fahim:au 3. #1 AND #2 4. #1 AND #2 AND [2000-2019]/py
	Medline (12th April 2019)	1. TX ATHENA 2. AU Lee OR AU Fahim 3. S1 AND S2
	PsycINFO (12th April 2019)	1. TX ATHENA 2. AU Lee OR AU Fahim 3. S1 AND S2 4. Limiters - Publication Year: 2000-2018
	SCOPUS (12th April 2019)	(ALL (Athena)) AND ((AUTHOR-NAME(Lee) OR AUTHOR-NAME (Fahim))) AND (LIMITE-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO

		(PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009) OR LIMIT-TO (PUBYEAR, 2008) OR LIMIT-TO (PUBYEAR, 2007) OR LIMIT-TO (PUBYEAR, 2006) OR LIMIT-TO (PUBYEAR, 2005) OR LIMIT-TO (PUBYEAR, 2004) OR LIMIT-TO (PUBYEAR, 2003) OR LIMIT-TO (PUBYEAR, 2002) OR LIMIT-TO (PUBYEAR, 2001) OR LIMIT-TO (PUBYEAR, 2000))
MAPS (Fanning et al., 2017)	CINAHL (11th April 2019)	1. TX MAPS OR TX “Multiphase activity promotion study” 2. AU Fanning OR AF Illinois OR AF Virginia 3. S1 AND S2 4. Limiters – Published Date: 20170101-20191231
	EMBASE (11th April 2019)	1. maps OR ‘multiphase activity promotion study’ 2. fanning:au OR ‘university of illinois’.ff OR ‘university of virginia’:ff 3. (#1 AND #2) AND [2000-2019]/py
	Medline (12th April 2019)	1. TX MAPS OR TX “Multiphase activity promotion study” 2. AU Fanning OR AF Illinois OR AF Virginia 3. S1 AND S2 4. Limiters – Date of Publication 20000101-20191231
	PsycINFO (12th April 2019)	1. TX MAPS OR TX “Multiphase activity promotion study” 2. AU Fanning OR AF Illinois OR AF Virginia 3. S1 AND S2 4. Publication Year: 2000-2019
	SCOPUS (12th April 2019)	(ALL (maps)) OR ALL (“Multiphase activity promotion study”))) AND ((AUTHOR-NAME(Fanning) OR AFFIL (illinois) OR AFFIL (virginia))) AND(LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR

		LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009) OR LIMIT-TO (PUBYEAR, 2008) OR LIMIT-TO (PUBYEAR, 2007) OR LIMIT-TO (PUBYEAR, 2006) OR LIMIT-TO (PUBYEAR, 2005) OR LIMIT-TO (PUBYEAR, 2004) OR LIMIT-TO (PUBYEAR, 2003) OR LIMIT-TO (PUBYEAR, 2002) OR LIMIT-TO (PUBYEAR, 2001) OR LIMIT-TO (PUBYEAR, 2000))
Unnamed (Fernández et al., 2013)	CINAHL (3rd May 2019)	1. AU Fernandez AND AF Catalunya 2. Limiters – Published Data: 200000101-20191231
	EMBASE (3rd May 2019)	Fernandez:au AND catalunya:ff AND [2000-2019]/py
	Medline (3rd May 2019)	1. AU Fernandez AND AF Catalunya 2. Limiters – Published Data: 200000101-20191231
	PsycINFO (3rd May 2019)	1. AU Fernandez AND AF Catalunya 2. Limiters – Published Data: 20000101-20191231
	SCOPUS (3rd May 2019)	(AUTHOR-NAME (Fernandez) AND AFFIL (catalunya)) AND PUBYEAR>1999
Haptivity (Forster et al., 2017)	CINAHL (11th April 2019)	TX Haptivity
	EMBASE (12th April 2019)	Haptivity
	Medline (12th April 2019)	TX Haptivity
	PsycINFO (12th April 2019)	TX Haptivity
	SCOPUS (12th April 2019)	ALL (Haptivity)
Up	CINAHL (11th April 2019)	TX “Jawbone Up app” OR TX “Up app”
	EMBASE (12th April 2019)	‘jawbone up app’ OR ‘up app’
	Medline (12th	TX “Jawbone Up app” OR TX “Up

	April 2019)	app”
	PsycINFO (12th April 2019)	TX “Jawbone Up app” OR TX “Up app”
	SCOPUS (12th April 2019)	(ALL(“jawbone up app”) OR ALL (“up app”))
Ngala (Hearn et al., 2014)	CINAHL (11th April 2019)	TX “Ngala healthy you healthy baby” OR TX HYHB OR TX “healthy you healthy baby”
	EMBASE (12th April 2019)	‘Ngala healthy you healthy baby’ OR HYHB OR ‘healthy you healthy baby’
	Medline (12th April 2019)	1. TX “Ngala healthy you healthy baby” OR TX HYHB OR TX “healthy you healthy baby” 2. Limiters – Date of Publication: 20000101-20191231
	PsycINFO (12th April 2019)	TX “Ngala healthy you healthy baby” OR TX HYHB OR TX “healthy you healthy baby”
	SCOPUS (12th April 2019)	(ALL(“Ngala healthy you healthy baby”) OR ALL (hyhb) OR ALL (“healthy you healthy baby”))
iN Touch (Kim et al., 2015)	CINAHL (12th April 2019)	1. TX “iN Touch” OR TX TheCarrot 2. AU Katherine Kim OR TX Trauner OR AF Davis OR AF San Francisco 3. S1 AND S2 4. Limiters – Published Date: 20000101-20191231
	EMBASE (12th April 2019)	1. ‘in touch’ OR thecarrot 2. Kim:au OR trainer OR davis:ff OR ‘san francisco’:ff 3. (#1 AND #2) AND [2000-2019]/py
	Medline (12th April 2019)	1. TX “iN Touch” OR TX TheCarrot 2. AU Katherine Kim OR TX Trauner OR AF Davis OR AF San Francisco 3. S1 AND S2 4. Limiters – Published Date: 20000101-20191231
	PsycINFO (12th April 2019)	1. TX “iN Touch” OR TX TheCarrot 2. AU Katherine Kim OR TX Trauner OR AF Davis OR AF San Francisco 3. S1 AND S2 4. Limiters – Publication Year: 2000-2019

	SCOPUS (12th April 2019)	(ALL (“iN Touch)) OR ALL (thecarrot))) AND ((AUTHOR-NAME(kim) OR ALL (kim) OR ALL (trainer) OR AFFIL (davis) OR AFFIL (“San Francisco”))) AND (LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009) OR LIMIT-TO (PUBYEAR, 2008) OR LIMIT-TO (PUBYEAR, 2007) OR LIMIT-TO (PUBYEAR, 2006) OR LIMIT-TO (PUBYEAR, 2005) OR LIMIT-TO (PUBYEAR, 2004) OR LIMIT-TO (PUBYEAR, 2003) OR LIMIT-TO (PUBYEAR, 2002) OR LIMIT-TO (PUBYEAR, 2001) OR LIMIT-TO (PUBYEAR, 2000))
SIGMA (Podina et al., 2017)	CINAHL (11th April 2019)	1. TX SIGMA OR TX SIGMAe OR TX SIGMAi OR TX (“Self-help integrated and gamified mobile application”) 2. AU Podina OR AF Bucharest OR AF Babes-bolyai 3. S1 AND S2
	EMBASE (12th April 2019)	1. sigma OR sigmae OR sigmai OR ‘self-help integrated and gamified mobile application' 2. podina:au OR Bucharest:ff OR ‘babes bolyai’:ff 3. (#1 AND #2) AND [2000-2019]/py
	Medline (12th April 2019)	1. TX SIGMA OR TX SIGMAe OR TX SIGMAi OR TX (“Self-help integrated and gamified mobile application”) 2. AU Podina OR AF Bucharest OR AF Babes-bolyai 3. S1 AND S2 4. Limiters – Date of Publication: 20000101-20191231
	PsycINFO (12th April 2019)	1. TX SIGMA OR TX SIGMAe OR TX SIGMAi OR TX (“Self-help integrated and gamified mobile

		application") 2. AU Podina OR AF Bucharest OR AF Babes-bolyai 3. S1 AND S2 4. Limiters – Publication Year: 2000-2019
	SCOPUS (12th April 2019)	(ALL (sigma) OR ALL (sigmae) OR ALL (sigmai) OR ALL ("self-help integrated and gamified mobile application"))) AND ((AUTHOR-NAME (podina) AFFIL (bucharest) OR AFFIL (babes-bolyai))) AND (LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009) OR LIMIT-TO (PUBYEAR, 2008) OR LIMIT-TO (PUBYEAR, 2007) OR LIMIT-TO (PUBYEAR, 2006) OR LIMIT-TO (PUBYEAR, 2005) OR LIMIT-TO (PUBYEAR, 2004) OR LIMIT-TO (PUBYEAR, 2003) OR LIMIT-TO (PUBYEAR, 2002) OR LIMIT-TO (PUBYEAR, 2001) OR LIMIT-TO (PUBYEAR, 2000))

Appendix 11 Ethics checklist application for public app evaluation

UNIVERSITY of BRADFORD

EC ....

(For Office Use Only)

APPLICANT'S ETHICS CHECKLIST
YOU MUST COMPLETE THIS FORM IF YOUR WORK INVOLVES HUMAN PARTICIPANTS, THEIR TISSUE OR THEIR DATA

This checklist is designed to help you to decide whether or not ethics approval is required and, if required, to decide on the appropriate ethics review procedure – please read Annex 1 on page 5 before you complete this form

Please Note:

a) All questions on this checklist should be completed. Once completed, submit the checklist to ethics@bradford.ac.uk.

b) Contact details (email address) should be given for PI or PS and student (if applicable).

c) In the case of Student projects, Supervisors should read and sign this checklist (in the correct box – EITHER/OR – not both boxes) BEFORE it is submitted to the Ethics Administrator for sign off by the Chair of the Research Ethics Panel.

d) Guidance on the 2 different ethics review procedures that together make up the University's Ethics Review System (i.e. 'University' and 'NHS') is available on the University Ethics website.

e) It is expected in the first instance, that if your project will involve human tissue/biological fluids you must contact the UoB Designated Individual for the HTA licence, Dr Mojgan Najafzadeh (M.Najafzadeh1@bradford.ac.uk or on 01274 236290) for audit purposes.
or Joanne Mullarkey or Wayne Burrill in Ethical Tissue on J.Mullarkey@bradford.ac.uk, W.Burrill@bradford.ac.uk or on 01274 235818 for audit purposes and to check if it can be sourced through them.

f) **If this Checklist is NOT correctly completed, it will be returned to you unauthorised.**

g) Please Note: If amendments are required after review, please submit a list of the changes made. Alter the original documents using 'track changes' or highlight what has been amended and re-submit the amended paperwork. This will allow the reviewers/Chair to easily identify what has been amended and improve the speed of the process.

–

Project Title: Quality and characteristics of publicly available apps that promote physical activity by providing feedback on mood

Name of Principal Investigator / Principal Supervisor: Neil Small

Contact Details – email address: N.A.Small@bradford.ac.uk

Department/School Faculty of Health Studies

Name of Student (if applicable)**: Laura Lamming**

Contact Details – email address L.Lamming1@bradford.ac.uk

Please indicate which Panel you think should review your checklist/application:

Humanities, Social and Health Sciences Research Ethics Panel **(HSHS)**. ☒

Biomedical, Natural, Physical and Health Sciences Research Ethics Panel **(BNPHS)**. ☐

Has the Principal Investigator / Principal Supervisor attended appropriate ethics training? Yes☒ No ☐

In case of student projects, please tick to confirm the following:

Post-Graduate Taught programme ☐ Under-Graduate Taught programme ☐ Post-Graduate Research ☒

Please indicate any deadlines this application must meet in order for the student to complete their research within course boundaries: 14th April 2019

Has the student attended appropriate ethics training? Yes☒ No ☐

Please give summary of project (max 150 words**):**

Background: Despite the documented health benefits, the UK population is inactive. One barrier may be the distal or unobservable nature of many of the benefits of physical activity (e.g. weight loss, reduced blood pressure). Raising awareness of the acute (immediate) benefits of physical activity (increase in positive mood/wellbeing), may promote behaviour change. Apps offer the opportunity to monitor both real-time physical activity and mood and provide intervention. Their quality is still in question due to their rapid proliferation and limited involvement from behaviour change experts. This study aims to assess the current evidence (characteristics and quality) for publicly available smartphone applications (apps) that capture and provide feedback on mood and physical activity levels (PA), with the goal of increasing subsequent physical activity (PA).

Aims and objectives: What is the current evidence for using feedback on immediate positive affect to promote PA in adults via a publicly available smartphone app?

a. Do any exist b. What are their characteristics in terms of method of feedback, target physical activity, target population, platform and cost c. What is their quality in terms of theoretical underpinnings, use of appropriate behaviour change techniques, method of physical activity measurement, development processes, content credibility, functionality, aesthetics, currency, security, potential impact, engagement, and usage and user perceptions. A systematic evaluation of publicly available apps in the app stores will be conducted to identify apps that include feedback on affect/mood/emotion for adults. Quality and characteristics will be captured and assessed. **Methods:** The top 100 paid and free apps from the 'Health & Fitness' category of the Apple App and Google Play stores will be screened for whether or not they are promoting physical activity in healthy adults and providing feedback on user mood. Eligible apps will then be downloaded onto a study device and trialled for 48 hours to determine full functionality and content. The first 50 user reviews for eligible apps will be exported and anonymised. The apps, their descriptions and user reviews will be assessed for quality based on 13 indicators using a piloted data extraction proforma and standardised tools including the Mobile App Rating Scale and Theoretical Coding Scheme.

Q1a	**Is the proposed work a research project, i.e.** an 'investigation undertaken in order to gain knowledge and understanding'? (This includes work of educational value designed to improve understanding of the research process.)? **This may include surveys of current views, attitudes and opinions; especially where the information gathered is personal or sensitive.** If you answer 'Yes' to Q1a ethical approval may be required, move to Q2. If you answer 'No' to Q1a then a research ethics review is not usually required; please move to question 1b and 1c. ***Note**: there may be occasions where a project is not defined as research but still raises ethical issues – please submit for review if this is the case.*	Yes ☒ No ☐
Q1b	**Is the proposed project an audit? This includes the organisation and analysis of data that already exists, as well as the comparison of service delivery with established standards (or standards in development).**	Yes ☒ No ☐
Q1c	**Is the proposed project service evaluation? E.g., does it investigate the functioning of existing processes or evaluate modifications to existing processes? This may include short anonymous surveys of user satisfaction that do not include sensitive or personal information.** A more detailed definition of Research, Audit and Service Evaluation is available here. **If you answered 'Yes' to Q1b, please explain below how humans are involved?**	Yes ☐ No ☒

Q2	**Will the project involve the <u>NHS</u>?** If you answer 'No' to Q2 move on to Q3 If you answer 'Yes' to Q2 ethical approval may be required by NHS Research Ethics Committee (REC), please use the Decision Tool to find out what approvals you will need. Please submit your IRAS checklist and application to the University via nhs-ethics@bradford.ac.uk for review. You will then be informed after review that you have approval to submit it via the IRAS Portal for HRA approval; the process can be found here.	Yes ☐ No ☒
Q3	**Will the project involve any of the following in the UK:** ▪ Testing a medicinal product ▪ Investigating a Class II or higher medical device* (see https://www.gov.uk/government/publications/managing-medical-devices for further details). ▪ Taking samples of human biological material (e.g. blood, tissue) ▪ Prisoners or others in custodial care (e.g. young offenders) as participants ▪ Adults with mental incapacity as participants ▪ Other vulnerable groups (e.g. vulnerable children) as participants If you answer 'Yes' to Q3 ethical approval will <u>*usually*</u> be required through a Research Ethics Panel, Ethical Tissue or NHS Research Ethics Committee (REC), or where the project includes participants that need approval under the Mental Capacity Act, approval will be required by the Social Care REC. If you wish to source material from Ethical Tissue at the University, they can be contacted on 01274 235897 or visit https://www.bradford.ac.uk/business/ethical-tissue/ If your work involves a medical device, please state it's Class according to the Medical Devices Directive (93/42/EEC) (see https://www.gov.uk/government/publications/managing-medical-devices *for further details). *If you answer 'No' to Q3 move on to Q4*	Yes ☐ No ☒
Q4	**Will the project involve human participants and/or human data** *If you ticked 'Yes' please give details of the protocols for contacts with participants including:* • *Details of the methodology used (for example interview or observation)* • *Arrangements for identifying and contacting participants* • *Arrangements for ensuring informed consent* • *Arrangements for ensuring participants ability to withdraw* *Arrangements for accessing, storing and destroying the data*	Yes ☒ No ☐

	Human data will be collected in the form of user reviews posted to the app store websites for eligible apps. The App Store functionality means that reviewers cannot be contacted to gain consent to use their reviews. The App Store guidance for reviewers asks that no personally identifiable information is posted within a review (e.g. https://play.google.com/about/comment-posting-policy.html?hl=en-GB) In accordance with the Association of Internet Researchers ethics recommendations, protection of individual user privacy was considered based on the acknowledged publicity of the App Stores. App stores and their review functions are clearly in the public domain as reviews can be accessed without any form of registration or login. Therefore, as the reviews are considered public and reviewers cannot be contacted, no consent process can take place or is deemed necessary. In addition, apps targeting adults are sought therefore user reviews should not include reviews from minors. The nature of the data is not sensitive, constituting reviews of physical activity app functionality and preferences. However, user names or handles and any identifiable information will be removed/anonymised and data will be stored in that state. Analysis may result in use of quotes, but again, if necessary, these will be anonymised before reporting. • **If your project involves NHS staff as participants, please complete this checklist providing the details listed above at the end of the checklist so that a decision can be made as to whether you need to complete an IRAS Checklist and IRAS Application form.** • If an application is required it will be under the IRAS system. Please submit your IRAS checklist and application to the University via nhs-ethics@bradford.ac.uk for review before submitting it via the IRAS Portal for HRA approval. The process can be found here.	
Q5a	**Will the research project involve <u>human tissue (but not requiring NHS approval see Q3)</u>?** If you answer 'Yes' to Q5 University ethical approval is required If your project will involve human tissue/biological fluids you must contact the UoB Designated Individual for the HTA licence, Dr Mojgan Najafzadeh (M.Najafzadeh1@bradford.ac.uk or on 01274 236290) for audit purposes. or Joanne Mullarkey or Wayne Burrill in Ethical Tissue on J.Mullarkey@bradford.ac.uk, or on 01274 235818 for audit purposes and to check if it can be sourced through them. You can also visit www.ethicaltissue.org	Yes ☐ No ☒
Q5b	If you answered 'Yes' to Q5a, is the human material over 100 years old and archaeological? If 'YES' please refer to the Biological Anthropology Research Centre (BARC)	Yes ☐

	guidelines at https://www.bradford.ac.uk/life-sciences/media/lifesciences/schooloflifesciences/ages/BARC_human_remains_policy.pdf	No ☐
If you answer 'No' to Q5 and have answered 'No' to Q2, Q3 and Q4 ethical approval is <u>not</u> required.		

PLEASE COMPLETE and SIGN *<u>ONE</u>* of the two boxes below

Please email the completed checklist form together with any supporting documents to the Ethics Administrator (ethics@bradford.ac.uk). *Please note* that we do *require a Principal Investigator's wet signature,* before we can have the application reviewed by the Research Ethics Panel this can be sent or delivered to RKTS, F.24, Richmond Building.

1. I have discussed this project with my student AND/OR
2. I confirm that there are <u>no ethical issues</u> requiring further consideration.

If you consider that ethical review is not required, please explain briefly why not, below:

Data being collected is already in the public domain. Data is not considered sensitive or contentious or harmful to users who submitted the reviews in any way and where appropriate will be anonymised and stored securely.

(Any subsequent changes to the nature of the project will require that the Panel are informed of all changes)

Signed by (Principal Investigator or Principal Supervisor (in case of student project)):
Signature:

N. A. Small.

Date: 15/3/19

PLEASE PRINT NAME: Neil Small

OR

I confirm that there are <u>ethical issues</u> requiring further consideration and will either:

1. Refer the proposal to Ethical Tissue, or,

2. Fill in and submit a full ethics application to be considered by the appropriate Research Ethics Panel, or,

3. Fill in a generic full application where generic approval is required (i.e., an application that will cover several sufficiently similar research projects with similar methodology or projects where the difference between the tests involved is not ethically relevant).

4. If your project involves NHS staff as participants and requires ethical approval, please provide details below:

5.

Name (Principal Investigator/Principal Supervisor):

Signature: ………………………………………………………….. Date: ……………………………..

PLEASE PRINT NAME ………………………………………………………………………………..

Annex 1

Ethical Scrutiny by a University Research Ethics Panel is not required if:

- The project is NOT a research project. There may be occasions where a project is not defined as research but still raises ethical issues – please submit for review if you think this is the case.

- The research project will only involve unlinked or aggregated human data which was collected and which was, at the time, subject to relevant research ethics panel approval. However, where this is the case the researcher should at least confirm this in an email to the Research Support Unit's Ethics Administrator so that the Ethics Administrator has a record and can inform the Chair of the appropriate Research Ethics Panel that the researcher plans to go ahead without ethics approval. The email should confirm that the research project does not require ethics approval because it only involves unlinked or aggregated data, which when originally obtained from people was obtained in accordance with the protocol as approved at the time by an appropriate research ethics panel. The email should also briefly explain how the researcher now plans to use the unlinked or aggregated data.

- The research is Public Domain Data:
  The Economic and Social Research Council's (ESRC) Research Ethics Framework states that ethics approval may not be required for data sets that exist in the public domain (e.g. datasets that are available from the Office for National Statistics or from the ESRC's Data Archive) so long as the appropriate permissions from individuals have already been obtained (i.e. informed consent) and where it is not possible to identify the individuals from the information provided. It must be remembered that public domain data is still covered by the laws of copyright.

- The research involves Simple Uncontentious Questionnaires:
  If a research project's only involvement with human subjects is a simple brief questionnaire with uncontroversial content it may not require ethical approval. It is the Principal Investigator or Principal Supervisor's responsibility to decide whether a project comes under this category and must indicate this at Q.4 on the checklist and attach the questionnaire document for information.

Guidance on supervisor and principal investigator sign off of uncontentious research

Audit and service evaluation are usually uncontentious, and guidance on how to differentiate between research, audit and service evaluation is given at: University Ethics website.

Even where a project is clearly research, as a supervisor or principal investigator, you can sign off simple, ethically uncontentious projects as not needing further ethical scrutiny. To do this, you should consider the level of risk to participants and researchers, the level of effort required by participants, the level of intrusion into participants' lives and the level of sensitivity of both the general subject matter and the information requested of participants. Basically, the lower these levels, the more likely the research is to be uncontentious and the more confident you should feel about signing off.

The following examples may help.

These studies can almost always be signed off by the supervisor or principal investigator:

- Brief questionnaires asking opinions about matters which are clearly not sensitive (attitudes to a product, beliefs about the usefulness of a course).
- Brief interviews about such topic.
- Observational studies about everyday behaviour in public places which involve no risk to subjects or the researcher.

But the following studies almost always need further scrutiny by a University Ethics Panel:

- Long questionnaires (these require considerable potential inconvenience to subjects).
- Long interviews
- Any questionnaires which ask subjects about intimate behaviours or issue likely to cause distress or would in other ways normally be regarded as contentious or sensitive (e.g. illegal activities, attitudes to abortion, capital punishment, immigration, euthanasia).
- Any interviews which examine these matters.
- Observational studies which involve intimate behaviours, behaviours which are not normally public or which might normally be considered contentious or sensitive (Activities of ethics committees, appointment committees, etc; professional consultations).

Naturally, this list is for illustration only, and should not be considered in any way exhaustive, permissive or prescriptive. For example, there are many categories of research not mentioned here which would definitely require ethics approval (e.g. treatment research). Rather the list demonstrates the issue of proportionality. Thus, even though the method may be the same for activities requiring and not requiring further scrutiny, the *content* in some way distinguishes between the two categories.

At the same time, there is obviously some middle ground. Are ethics committees not public? Is what is discussed so sensitive that the proposal needs further scrutiny? What about asking people about their views on the actions of senior members of staff in their organisation? Probably, it is in these middle ground areas that further advice should be sought from a Panel Chair about whether the project can be signed off by the supervisor or principal investigator alone. Given that, in so doing, the supervisor or PI is attesting to the ethical probity of the study, it is usually best to err on the side of caution where there is uncertainty. Panel chairs are very happy to advise.

Dr Martin Brinkworth, Chair, Biomedical, Natural, Physical and Health Sciences Research Ethics Panel, m.h.brinkworth@bradford.ac.uk, ext. 3584

Dr Clare Beckett-Wrighton, Chair, Humanities, Social and Health Sciences Research Ethics Panel, c.beckett@bradford.ac.uk, ext. 3521)

Please submit this checklist to:
Ethics Administrator, RaIS,

F.24 Richmond Building

in hard copy or by email to

ethics@bradford.ac.uk

Appendix 12 Ethics approval for public app evaluation

Dear Laura and Neil,

Ethics Checklist: EC25643
Title: Quality and characteristics of publicly available apps that promote physical activity by providing feedback on mood

Your ethics submission, documents and amendments have now been reviewed by the Vice Chair of the Research Ethics Panel.

I am pleased to inform you that the Vice Chair has confirmed approval of this study with no further ethical scrutiny required.

NOTE that this approval is for this study only.
Should there be any changes to this study, you must inform ethics@bradford.ac.uk.
Once your changes have been reviewed and you have approval to proceed, only then can you recommence the study.
Failure to do so will render your original approval invalid and withdrawn.

Please add a sentence onto any material you share with participants confirming that ethics approval has been granted by the Chair of the Humanities, Social and Health Sciences Research Ethics Panel at the University of Bradford on 03/04/19.

Many thanks
Best wishes
Deborah

Deborah Hodgson
Research and Innovation Administrator
Research and Innovation Services (RaIS), F.24
Ext: 3196

Appendix 13 Title screening guidelines for public app evaluation

PICOS and guidelines for commercial app title screening

Evaluation of public apps for physical activity promotion

The spreadsheets have hyperlinks to the apps in App Annie– click on those to view the app descriptions and screenshots. Review BOTH for the purposes of this screening. Screenshots can be enlarged by clicking on them. In order to access App Annie you will need this login:

Email: XXXX

PW: XXXX

To view Android screenshots you may need to view the app in the app store directly as the images are tiny in some instances. Some apps also have videos, these will not play in App Annie.

Responses

Responses in the spreadsheet, including final eligibility decision, will turn red, green or grey depending on your answers. The colours should help indicate whether the paper should be included (downloaded), or excluded or if you're still unsure whether it needs to be downloaded – see item 6 below for details.

Colour	**Meaning**
Red	Criteria has definitely not been met and therefore paper will be excluded
Green	Criteria has been met and app will be downloaded
Grey	Unclear/insufficient information to be sure that criteria has been met – needs downloading

1. **Does it <u>promote</u> physical activity in anyway and/or its uptake, maintenance, promotion OR reducing sedentary behaviour either alone or as part of a lifestyle app?**

Response	**Possible reasons**
Y (yes)	Clearly looking at promoting some sort of physical activity (all variations on the term relevant) for the purposes of getting fit/being physically healthy/just being active/stronger/more flexible. The app is looking to change/influence/explore multiple behaviours that INCLUDE promoting physical activity for the

	sake of weight loss or general wellbeing/lifestyle change. I.e. where physical activity is not being used for fitness increasing/physical activity promotion purposes then exclude for example where exercise/steps are being tracked for weight loss purposes but are not being promoted e.g. calorie counters. BUT: if the calorie counter (or otherwise) appears to be actively promoting exercise (for example via exercise programmes or encouraging text or reward badges – NOT JUST TIPS, see 'No') to allow calorie consumption or just on its own, then respond YES. Can include apps that support an exercise programme for purposes of weight loss/toning/strength building/getting fit.
N (no)	Where physical activity tips alone are being used as part of an app that focuses on something other than general lifestyle improvement e.g. focus is on pregnancy alone. Exercise IS NOT being promoted in any way e.g. calorie counters. BUT, if the calorie counter appears to be actively promoting exercise to allow calorie consumption, then respond YES Looking at rehabilitation or bringing people back to a normal level of functioning. Looking at physical functioning in general or balance and falls. Any therapeutic physical activity/exercise or intervention e.g. using it to reduce anxiety as part of a meditation app or promotes specific exercises to improve sexual performance. Where physical activity is not mentioned at all and it seems highly unlikely that it's incorporated. E.g. where other behaviours not including physical activity are examined e.g. diet, smoking, alcohol consumption etc. Purely an interval timer or route tracker/map provider, or used to convert wearable data/app data between platforms. IF NO – STOP SCREENING NOW
U (unsure)	Any apps where you're not sure if PA is being promoted in its own right.

2. **Are the participants/target group for change free-living, condition-free, adults?**

Response	**Possible reasons**
Y (yes)	Makes no reference to a specialist population, then assume general population, otherwise:

	18 years or older – assume adults if referred to as ‘employees’ or ‘workers’ or ‘college/university students’ or ‘veterans’. AND Live in the community. AND Suffer from no condition/disease. HOWEVER, for the purposes of inclusivity and potential for healthy control groups the following can be included for now: • obese/overweight individuals
N (no)	Children/teenagers under the age of 18, e.g. high school students. Prisoners, hospitalised patients, those in assisted living. Those with a condition or disease including mental or physical health conditions. Those who have recently had a disease e.g. ‘post stroke’ or ‘cancer survivors’ or ‘recently had gestational diabetes’.
U (unsure)	Pregnant women. Those who are pre-disease e.g. prediabetic, prehypertensive. Those described as ‘at risk’ of a condition. Youth/young adults (unless age confirmed). Any other instance where it’s unclear what participant status is.

3. Is it in English?

Response	**Possible reasons**
Y (yes)	Title and description and app pics depict English text – can also have other languages depicted/used, as long as an English translation is available as a description AND an English version of the app is available
N (no)	Title, app or description not in English
U (unsure)	If there’s an English version that may be available

4. **Affect**

Response	Possible reasons
Y (yes)	Mention of mood/stress/affect/emotion is made in description in some way that may suggest it is being monitored within the app OR Screenshots suggest mood is captured e.g. smiley-face responses, "I feel…." Scales etc.
N (no)	No mention or visual depiction of mood observed
U (unsure)	Mood is mentioned/visually depicted in a way that makes it unclear if it will definitely be captured

5. **Include/Exclude/Unsure**

Response	Possible reasons
I (include)	All answers were green OR All answers except 4. Affect, were green and 4.Affect was grey
E (exclude)	At least one answer OTHER THAN 4.Affect was red OR - Purely a class booking app e.g. allows viewing timetables and booking classes only - Only functions to display wearable device information/requires a wearable to work AT ALL, as we cannot access wearables - Merely provides calorie information for different activities - Membership of some sort is required to access functionality/features
U (unsure)	If Answers 1-3 included at least one grey AND/OR 4.Affect was red. e.g. 1-3 could all be green but 4 is red OR e.g. 1 is grey, and the rest of items 2-4 are green

6. **Notes and other reasons for exclusion**

Make a note if:

- Duplicate
- Want to clarify any responses you've made

- The reasons why it was excluded if it wasn’t due to items 1-4 (see exclude item in 5, or any other reasons you can think of)
- The app is no longer available in the app store
- Any other concerns you might have

Appendix 14 Search strategies for eligible app studies for public app evaluation

App and reference	Database and date search conducted	Search strategy
Clue	CINAHL (21st May 2019)	1. TX "Period Tracker Clue Period & Ovulation Tracker" OR TX BioWink OR TX Clue 2. TX app OR TX application 3. S1 AND S2 4. Limiters – Published Date: 20000101-20190631
	EMBASE (21st May 2019)	1. 'Period tracker clue period & ovulation tracker' OR clue OR Biowink 2. App OR application 3. #1 AND #2 [2000-2019]/py
	Medline (21st May 2019)	1. TX "Period Tracker Clue Period & Ovulation Tracker" OR TX BioWink OR TX Clue 2. TX app OR TX application 3. S1 AND S2 4. Limiters – Published Date: 20000101-20190631
	PsycINFO (21st May 2019)	1. TX "Period Tracker Clue Period & Ovulation Tracker" OR TX BioWink OR TX Clue 2. TX app OR TX application 3. S1 AND S2 4. Limiters – Publication Year: 2000-2019
	SCOPUS (21st May 2019)	(ALL (Clue) OR ALL (biowink) OR ALL ("period tracker clue period & ovulation tracker"))) AND ((ALL (app) OR ALL (application))) AND (LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010) OR LIMIT-TO (PUBYEAR, 2009) OR LIMIT-TO (PUBYEAR, 2008) OR LIMIT-TO (PUBYEAR, 2007) OR LIMIT-TO

		(PUBYEAR, 2006) OR LIMIT-TO (PUBYEAR, 2005) OR LIMIT-TO (PUBYEAR, 2004) OR LIMIT-TO (PUBYEAR, 2003) OR LIMIT-TO (PUBYEAR, 2002) OR LIMIT-TO (PUBYEAR, 2001) OR LIMIT-TO (PUBYEAR, 2000))
Fitbit	CINAHL (31st May 2019)	1. TX fitbit 2. TX app OR TX application 3. S1 AND S2
	EMBASE (3rd June 2019)	1. Fitbit 2. App OR application 3. #1 AND #2
	Medline (31st May 2019)	1. TX fitbit 2. TX app OR TX application 3. S1 AND S2
	PsycINFO (3rd June 2019)	1. TX fitbit 2. TX app OR TX application 3. S1 AND S2
	SCOPUS (3rd June 2019)	((ALL(app) OR ALL (application))) AND (ALL(fitbit))
Keep	CINAHL (3rd June 2019)	1. TX “keep fitness” OR TX “keep workout” OR TX “keep trainer” 2. TX app or TX application 3. S1 AND S2
	EMBASE (3rd June 2019)	5. ‘keep fitness’ OR ‘keep workout’ OR ‘keep trainer’ 6. App OR application 7. #1 AND #2
	Medline (3rd June 2019)	1. TX “keep fitness” OR TX “keep workout” OR TX “keep trainer” 2. TX app or TX application 3. S1 AND S2
	PsycINFO (3rd June 2019)	5. TX “keep fitness” OR TX “keep workout” OR TX “keep trainer” 6. TX app or TX application 7. S1 AND S2
	SCOPUS (3rd June 2019)	((ALL (“keep trainer”) OR ALL (“keep workout”) OR ALL (“keep fitness”))) AND ((ALL (app) OR ALL (application)))
On You Couch to 5k	CINAHL (21st May 2019)	5. TX “one you couch to 5k” OR TX “public health England”OR TX “couch to 5k” 6. TX app OR TX application 7. S1 AND S2
	EMBASE (21st May 2019)	4. ‘one you couch to 5k’OR ‘public health england’ OR ‘couch to 5k’ 5. App OR application

		6. #1 AND #2 AND [2000-2019]/py
	Medline (21st May 2019)	5. TX “one you couch to 5k” OR TX “public health England”OR TX “couch to 5k” 6. TX app OR TX application 7. S1 AND S2
	PsycINFO (21st May 2019)	5. TX “one you couch to 5k” OR TX “public health England”OR TX “couch to 5k” 6. TX app OR TX application 7. S1 AND S2
	SCOPUS (21st May 2019)	((ALL (app) OR ALL (application))) AND ((ALL(“one you couch to 5k”) OR ALL (“public health England”) OR ALL (“couch to 5k”)))
Runkeeper	CINAHL (31st May 2019)	3. TX runkeeper 4. TX app OR TX application 5. S1 AND S2
	EMBASE (31st May 2019)	1. Runkeeper 2. App OR application 3. #1 AND #2
	Medline (31st May 2019)	3. TX runkeeper 4. TX app OR TX application 5. S1 AND S2
	PsycINFO (31st May 2019)	3. TX runkeeper 4. TX app OR TX application 5. S1 AND S2
	SCOPUS (31st May 2019)	((ALL (app) OR ALL (application))) AND (ALL (Runkeeper))
Runtastic Heart Rate PRO	CINAHL (3rd June 2019)	1. TX “runtastic heart rate pro” OR TX “Runtastic heart rate” 2. TX app OR TX application 3. S1 AND S2
	EMBASE (3rd June 2019)	1. ‘runtastic heart rate pro’ OR ‘runtastic heart rate’ 2. App OR application 3. #1 AND #2
	Medline (3rd June 2019)	1. TX “runtastic heart rate pro” OR TX “Runtastic heart rate” 2. TX app OR TX application 3. S1 AND S2
	PsycINFO (3rd June 2019)	1. TX “runtastic heart rate pro” OR TX “Runtastic heart rate” 2. TX app OR TX application 3. S1 AND S2
	SCOPUS (3rd June 2019)	((ALL (app) OR ALL (application))) AND ((ALL (“Runtastic heart rate pro”) OR ALL (“runtastic heart rate”)))
Runtastic PRO	CINAHL (3rd	1. TX “runtastic pro”

	June 2019)	2. TX app OR TX application 3. S1 AND S2
	EMBASE (3rd June 2019)	1. 'runtastic pro' 2. App OR application 3. #1 AND #2
	Medline (3rd June 2019)	1. TX "runtastic pro" 2. TX app OR TX application 3. S1 AND S2
	PsycINFO (3rd June 2019)	1. TX "runtastic pro" 2. TX app OR TX application 3. S1 AND S2
	SCOPUS (3rd June 2019)	((ALL (app) OR ALL (application))) AND (ALL ("Runtastic pro"))
Runtastic	CINAHL (30th May 2019)	1. TX Runtastic OR "runtastic running" 2. TX app OR TX application 3. S1 AND S2
	EMBASE (30th May 2019)	4. Runtastic OR 'runtastic running' 5. App OR application 6. #1 AND #2
	Medline (30th May 2019)	1. TX Runtastic OR "runtastic running" 2. TX app OR TX application 3. S1 AND S2
	PsycINFO (30th May 2019)	1. TX Runtastic OR "runtastic running" 2. TX app OR TX application 3. S1 AND S2
	SCOPUS (30th May 2019)	((ALL (app) OR ALL (application))) AND ((ALL (Runtastic) OR ALL ("Runtastic running")))
Samsung Health	CINAHL (31st May 2019)	5. TX "Samsung health" 6. TX app OR TX application 7. S1 AND S2 8. Limiters – Published Date: 20000101-20191231
	EMBASE (31st May 2019)	4. 'samsung health' 5. App OR application 6. #1 AND #2
	Medline (31st May 2019)	5. TX "Samsung health" 6. TX app OR TX application 7. S1 AND S2 8. Limiters – Published Date: 20000101-20191231
	PsycINFO (31st May 2019)	5. TX "Samsung health" 6. TX app OR TX application 7. S1 AND S2 8. Limiters – Published Date: 20000101-20191231

	SCOPUS (31st May 2019)	((ALL (app) OR ALL (application))) AND (ALL (“Samsung health”))

Appendix 15 Public app evaluation data extraction guidelines

Data extraction and quality assessment guidelines

Systematic Public App Evaluation

General guidelines and information

Research questions

The purpose of this data extraction form is to collect data from eligible literature to answer the following research questions:

1. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change?
2. What are the characteristics and content of physical activity apps that include feedback on immediate affect?
3. What is the quality of these physical activity apps, that include feedback on immediate affect?

Data extraction form and documents

The data extraction form consists of six excel sheets that allow for data extraction as well as quality assessment. Tabs include: Data Extraction, QA_TCS, QA_PA BCTs, QA_MARS and QA_PA measure reliability. This document will be divided into different sections per tab.

You will have been provided with:

- the app on an appropriate device
- a link to the app in the app store
- other associated papers/website links that provide evidence for the app if available
- the original App Annie ranking information for the app that provides details of star ratings, number of ratings and release dates
- a link to the App Annie page for the app

Where these are available ALL should be used to perform the data extraction. HOWEVER if the app website discusses MULTIPLE apps, only refer to information that is GENERIC and therefore applicable to the app being tested AS WELL AS the other apps and information about the app being tested specifically.

You should ‘test’ the app over a 48 hour period to ensure you are aware of all functionality, including registering with linked communities and any other services that are free and specific to the app. For example, app communities can be linked to and additional training content downloaded/unlocked, but additional app upgrades or in-app purchases should not be made if they cost.

These guidelines include a number of appendices. Some are at the end of this document and some are in separate documents which you will have been provided with, due to their size. Please see the guidelines for where each appendix is located.

If any item is not reported, write: unreported, unless otherwise stated.

A: Data extraction tab

1. **Admin details**

Item	Explanation
1.1 Coder initials	Enter your initials
1.2 Date that extraction and 48 hour testing period for app started	Enter the date you started the data extraction – this should be the same date that you started the 48 hour testing of the app
1.3 Date app was downloaded	Enter the date that the app was downloaded to the test device
1.4 App name	Full app name
1.5 Extra evidence provided?	List what was provided: • App store link • App Annie ranking information • App Annie link • App website • Academic papers
1.6 Extra devices required?	Were wearables or other devices required to make the app work and could a wearable/device be linked to the app?
1.7 Additional notes	If you have any other concerns/thoughts about this section, put them here.

2. **Targeted users**

Item	Explanation
2.1 Diagnostic criteria	If any criteria were specified for users, give details e.g. sedentary as defined by a specific PA self-report measure and units
2.2 Age	User age range if specified/advised.
2.3 Sex	User sex if specified/advised
2.4 Socio-	User socio-demographics if specified/advised e.g.

demographics	education level, salary etc.
2.5 Ethnicity	User ethnicity if specified/advised
2.6 Additional notes	Any other details specific to app users that were used as potential inclusion/exclusion criteria that have not been reported above. If you have any other concerns/thoughts about this section, put them here.

3. App

Item	**Explanation**
3.1 Size of app	Size of app – in MB on phone
3.2 QA_Security and privacy	Provide details of any security and privacy features that app has including but not limited to: • Privacy protection • Presence and availability of security policies • Import/export practices • Use of login passwords • Encryption • Cloud back up • Adherence to Data Protection standards
3.3 Brief description of app, what it does and how	Can be copy and pasted from store but use judgement over whether or not it seems too brief/long and adapt
3.4 Is app tailored/personalised	Brief description of techniques or methods used to tailor/personalise app/app content and what is used to do so e.g. names, sedentary status, stage of change, user-selected activities/PA, user-provided data of some sort etc.
3.5 QA_Was PA measured/captured within the app or via a linked wearable/device?	Yes/No/Unreported If physical activity was ONLY captured in a way that the user did NOT receive information about their physical activity levels at all, then code NO. For example, code NO if: • Physical activity is not captured/measured using the app or by any other device/survey that is linked to the app Code YES if: • Physical activity is captured/measured using the app <u>and</u> users are able to see the results during the intervention • Physical activity is NOT measured/captured by the app, but is captured by some external device/questionnaire and users are given

	information on their physical activity levels as part of the intervention e.g. users may be provided with a separate pedometer which will show them their steps throughout the intervention period.
3.6 QA_How?	Use lists reported in Taylor 2014, pg 48-55, and Strath et al 2013 (appendix 1 & 2 as separate documents) Provide: • name of tool/device/wearable • make and model if appropriate • settings for tool/device as necessary • self-report or objective • just the name of validated questionnaire and key citation if unaltered • alterations to validated questionnaires • full details of any unvalidated/new tool/survey/questionnaire – if multiple questions, provide location in paper rather than copying them all out. If a smartphone specific app is used, name it. If the smartphone's inbuilt accelerometer/pedometer or other sensor is used, specify this so that a distinction can be made between wearables and device-based measures.
3.7 Targeting PA and/or other behaviours?	Does the app ONLY promote or does it promote other health behaviours too? If multiple behaviours, list them, otherwise just say PA
3.8 Type of PA targeted and list of specific activities if specified	List if it's a specific sport/type or if its lifestyle activity, or if the activity is 'user defined' and therefore it could have been a variety of unknown activities Provide list of promoted activities if available
3.9 QA_Presence of PA recommendations for adults or other (specify)	Specify any PA recommendations that are provided as part of the app – if none, say no.
3.10 What type of mood was captured?	Describe the type of mood that was captured – if it appears to be PA contingent mood i.e. mood that has been generated or is directly related to performing an activity e.g. it's being measured immediately after an activity or during an activity. Generic mood might be captured during the day without being specifically associated with an activity.

	Also detail if a particular type of mood was being captured (how stressed do you feel?) rather than mood in general (how do you feel?) e.g. stress levels, sadness, anxiety.
3.11 How was mood captured?	Describe how the app (or otherwise) captured participant mood including tool, units and frequency
3.12 How was mood processed and then fed back in the context of PA?	Describe how the app (or otherwise) processed (by researchers or automatically by app) and fed back participant mood in the context of PA including tool, units and frequency. If feedback on mood was provided by anything other than the app, describe how/by who and in what format and how it was delivered in the context of PA
3.13 How long was fed back mood data available for users?	Was the data available for a day, week, month, entire app usage duration?
3.14 QA_BCTs	Using BCT Taxonomy v1.0 (separate appendix 3), list the behaviour change techniques used in the app to the best of your ability. This includes those used to target OTHER behaviours, not just PA, in lifestyle apps. Provide: • An indication of the strength of your conviction regarding whether or not the BCT is present using the following '+' indicators and rules o Present + (BCT present in all probability but evidence unclear) o Present ++ (BCT present beyond all reasonably doubt) Developers may explicitly report the BCTs they used, but if more BCTs seem apparent then list them and note that the developers did not specify that they were part of the app. Please note the coding principles taken from the BCT online training: 1) Learning principle 1: Only code BCTs that are directly applied to the target behaviour(s) and population(s) 2) Learning principle 2: Do not infer the presence of a BCT 3) Learning principle 3: Take care distinguishing between BCTs that differ in terms of their behaviour change type (i.e.

	behaviour versus outcome) 4) Learning principle 4: Code technical terms and packages of BCTs that map onto BCTs in the taxonomy 5) Learning principle 5: All BCT definitions in the taxonomy include an action verb
3.15 QA_PA BCTs	Using the list of BCTs associated with PA (appendix 4 at end of this document), copy across those BCTs reported in 3.14 that match the list. If some are clearly being used to target DIFFERENT behaviours, make a note of that, but still report them.
3.16 QA_Involvement of users in development/pilot testing	Give details of if and how potential users were involved in the development/design/testing of the app. Potentially available in supporting evidence.
3.17 QA_Demographics of users involved in development/pilot testing	Give demographic details of said potential users including age, sex, ethnicity if provided OR specify if different to target population/sample population listed in ‘Participants using app’ section above. Potentially available in supporting evidence.
3.18 Was anything else used along with the app, to provide the PA and mood elements of the intervention? (e.g. wearable, website, counselling session)	Some apps may not be standalone but may have been supported by other things such as websites, wearables etc. – give details
3.19 Describe how these additional tools were used	Describe how they were used for the behaviours of interest – PA and affect/mood, capturing these and providing feedback on them.
3.20 Additional notes	If you have any other concerns/thoughts about this section, put them here.

4. Outcomes

Item	Explanation
4.1 QA_Was usage or compliance captured/assessed?	Yes/No/Unreported Not to be confused with engagement, usage or compliance refers to responses to app content such a required step-count submissions or response to prompts to complete questions. Can include rates of usage, number of time app opened, views of a certain app pages etc. Potentially available in supporting evidence.
4.2 QA_How? Including units of	Provide name of tool or units of measurement. Can include rates of usage.

measurement if applicable	Potentially available in supporting evidence.
4.3 QA_Was app quality measured/assessed?	Yes/No/Unreported This does not include usage, captured above. But can include measures of the following as defined in appendix 5 at the end of this document: • Acceptability/Participant perceptions • Aesthetics • Credible/trustworthy/appropriate or useful/essential information • Engagement • Functionality/Usability If terminology in paper is not the same as above, but definition seems to match, use paper terminology and mention which of the above definitions for the above terms you think it matches, for clarity. Or quality in general e.g. an overall quality score. Can also include general perceptions of quality for example via focus groups/piloting/interviews/user-testing/user reviews. Potentially available in supporting evidence.
4.4 QA_How? Including unit of measurement if applicable	Provide name of tool if appropriate (e.g. Mobile App Rating Scale) – self-report or objective and any alterations or settings for tool as necessary. If quality assessed more generically in terms of testing, report here. If unsure of relevance of measure, add and flag as unsure. Potentially available in supporting evidence.
4.5 Additional notes	If you have any other concerns/thoughts about this section, put them here.

5. Results

Item	Explanation
5.1 QA_Usage results	Results of usage assessment as reported in 4.1/4.2. Potentially available in supporting evidence.
5.2 QA_Quality measure results	Results of the quality assessment measure as reported in 4.3/4.4.

	Report findings both during development AND from final version of app if available and distinguish between the two. Potentially available in supporting evidence.
5.3 QA_Were quality issues addressed and if so, how?	Yes/No/Unreported Report findings both during development AND from final version of app if available and distinguish between the two. Describe briefly what changes were made based on results from quality assessment. Potentially available in supporting evidence.
5.4 Additional notes	If you have any other concerns/thoughts about this section, put them here.

6. Other

Item	**Explanation**
6.1 Any other comments/discussion/ conclusions relevant to review questions	Note any other details that you think are relevant to the review questions that haven't been captured elsewhere. **Research questions:** The purpose of this data extraction form is to collect data from eligible literature to answer the following research questions: 4. Are there any physical activity apps that include feedback on immediate affect (mood) to facilitate behaviour change? 5. What are the characteristics and content of physical activity apps that include feedback on immediate affect? 6. What is the quality of these physical activity apps, that include feedback on immediate affect?
6.2 References to other relevant studies/linked papers reporting on app further	Note references that may provide more information on the app or another potentially relevant app

B: QA_TCS: Theoretical Coding Scheme

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

TCS generic guidelines, adapted for thesis

The coding scheme that follows comprises 19 items, only 11 of which are captured as part of this data extraction/quality assessment. For each item, code as Yes or No. When 'Yes', state the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number]).

The coding scheme is based on the *explicit* use of theory. Consequently, even when theory-relevant information is implied but is not stated explicitly, the related items should be coded as 'no'.

Defining Terms

Please refer to the definitions (see below) during coding:

Theory (or Model)
'a set of interrelated concepts, definitions and propositions that present a *systematic* view of events or situations by specifying relations among variables, in order to *explain* or *predict* the events or situations' (Glanz & Rimer, 1995). Examples include: Theory of Planned Behaviour (TPB), Theory of Reasoned Action (TRA), Health Belief Model (HBM), stage of change/trans-theoretical model etc. Please refer to the '*Table of Theories*' below, but if the theory or construct is not represented in this table, refer to 'ABC of behaviour change theories' book (separate appendix 6). Try searching for the construct to see which theories it might relate to.

Theory-relevant construct
A construct (a key concept, excluding behaviour) within a theory/model upon which the intervention is based. Please refer to the '*Table of Theories*' below to identify whether a particular construct belongs to the specified theory. If the theory or construct is not represented in this table, refer to 'ABC of behaviour change theories' book (separate appendix 6). Try searching for the construct to see which theories it might relate to.

Predictors
Constructs that are not explicitly linked to a theory by the authors, but are targeted for intervention (as a means to change behaviour) because they predict behaviour. Predictors must only be coded if the author has presented evidence that the construct predicts/correlates with/causes behaviour. Predictors do not include actual behaviour, self-reported or otherwise (e.g. amount of time spent exercising), or biological factors (e.g., age, sex).

Intervention Technique
Strategy used to change behaviour, theory-relevant construct or predictor (e.g., providing information on consequences; prompting specific goal setting; prompting barrier identification; modelling the behaviour; planning social support). **NOTE: The TCS was developed before the first version of the BCT taxonomy was released, but BCTs can be coded in this section**.

Table of Theories

Theory	***Theory-relevant constructs***
Theory of Reasoned Action	1. **Attitudes**; 2. **Subjective Norms**; 3. **Intentions**
Theory of Planned Behaviour	1. **Attitudes**; 2. **Subjective Norms**; 3. **Perceived Behavioural Control**; 4. **Intentions**
Trans-Theoretical Model/Stages of Change	1. **Self-Efficacy** (person's confidence in performing a particular behaviour); 2. **Decisional Balance**
Social Cognitive Theory	1. **Self-Efficacy** (person's confidence in performing a particular behavior)/ **Behavioral capability** (Knowledge and skill to perform a given behavior); 2. **Action-Outcome Expectancies** (extent that one's actions are seen as instrumental for the outcome/values associated with outcomes)/**attitudes**; 3. **Barriers** (including changes to environment/emotional barriers or one's perceptions of them). The following constructs are also related to the theory (and subsequently should be listed when they are cited by the authors): *Behavioral capability:* Knowledge and skill to perform a given behavior; *Attitudes (outcome-expectancies):-* *Expectations:* Anticipatory outcomes of a behavior; *Expectancies:* The values that the person places on a given outcome, incentives; *Self-control:* Personal regulation of goal-directed behavior or performance *Goals?*

Health Belief Model	1. **Perceived Susceptibility**; 2. **Perceived Severity**; 3. **Perceived Benefits** (one's belief in the efficacy of the advised action to reduce risk or seriousness of impact); 4. **Perceived Barriers**; 5. **Cues to Action**; 6. **Self-Efficacy** (added by Rosenstock et al., 1988)
Protection Motivation Theory (PMT)	1. **Intention**; 2. **Perceived Behavioural Control**; 3. **Perceived Severity**; 4. **Perceived Vulnerability**; 5. **Response Efficacy**; 6. **Response Costs**; 7. **Fear** (now added to the model),
Rubicon Model / Model of Action Phases (Heckhausen, 1991; Gollwitzer, 1990)	1. **Motivation** (e.g., intention); 2. **Volition** (e.g., planning)

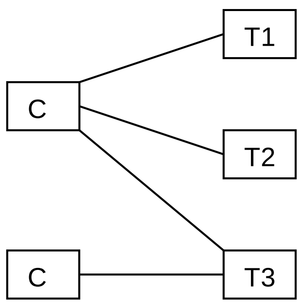
Notes:

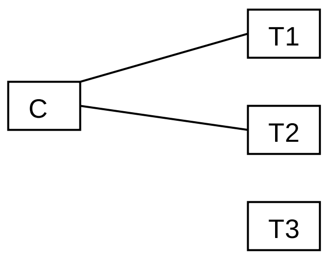
1. When more than one intervention is used (excluding the control group), the items should be coded separately for each intervention
2. Note for Stage of Change/Trans-Theoretical Model: While the construct 'stage of change' may either be used to select recipients of an intervention (item 4) or to tailor an intervention (item 6), 'stage of change' measures should not be coded for items 7-11. However, self-efficacy and decisional balance, both constructs within the TTM, should be coded throughout.
3. Where multiple behaviours are targeted by an app, if different theories/constructs/techniques are used to target different behaviours, report ALL and distinguish between those for different behaviours if possible.

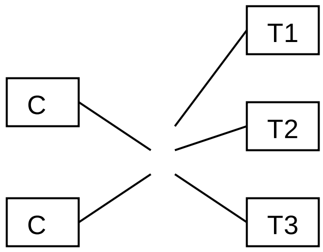
7. TCS item specific coding descriptions

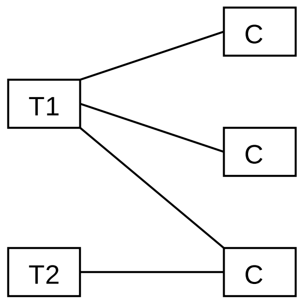
Item	Explanation
7.1a Theory/model of behaviour mentioned	Yes/No/Don't know Models/theories that specify relations among variables, in order to *explain* or *predict* behaviour (e.g., TPB, SCT, HBM) are mentioned, even if the intervention is not based on this theory
7.1b Theories and evidence location	List the theory/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.2a Targeted construct mentioned as predictor	Yes/No/Don't know

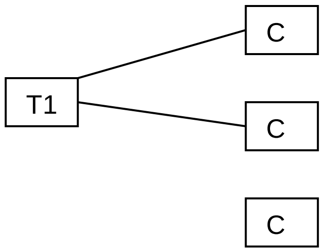
of behaviour	'Targeted' construct refers to a psychological construct that the study intervention is hypothesised to change. Evidence that the psychological construct relates to (correlates/predicts/causes) behaviour should be presented within the introduction or method (rather than the Discussion).
7.2b Construct and location of evidence that construct relates to behaviour	List the construct/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app). (These details are often found in the introduction section.)
7.2c Location that this predictor is targeted by the intervention	List the predictor/supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app). (These details are often found in the methods section.)
7.3a Intervention based on single theory?	Yes/No/Don't know The intervention is based on a single theory (rather than a combination of theories or theory + predictors)
7.3b Location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.4a Theory/ predictors used to select recipients for the intervention	Yes/No/Don't know Participants were screened/selected based on achieving a particular score/level on a theory-relevant construct/predictor
7.4b Construct (theory) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.4c Predictor and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.5a Theory/ predictors used to select/develop intervention techniques	Yes/No/Don't know The intervention is explicitly based on a theory or predictor or combination of theories or predictors (These details are often found in the Methods section and may be a replication of the findings for item 2.)
7.5bTheory and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.5c Predictor and	List the supporting evidence and its location, as

location	follows: ([insert page number], [insert paragraph number] or place reported in app).
7.6a Theory/ predictors used to tailor intervention techniques to recipients	Yes/No/Don’t know The intervention differs for different sub-groups that vary on a psychological construct (e.g., stage of change) or predictor at <u>baseline</u>
7.6b Construct and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.6c Predictor and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.7a All intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor	Yes/No/Don’t know Each intervention technique (T) is linked to at least one theory-relevant construct/predictor (C). Examples of meeting this criterion: - T1 & T2 are used to change C1; T3 is used to change C1 and C2. Techniques 1-3 are the only intervention techniques used (see DIAGRAM 1). - T1 used to change C1; T2 used to change C2; T3 used to change C3 **DIAGRAM 1: Example: All intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)**  In some cases, items 7, 8 and 9 may all be coded as ‘no’ (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa). If item 7 is coded ‘yes’ then items 8 and 9 must be c ‘no’. In instances where the name of the *construct* might overlap with the name of the technique (e.g., tips on *social support* were given; exercise

	benefits were discussed), code this as a direct link between technique and construct (i.e. they do not need to be written in the form 'tips on social support [technique] were used to target social support [construct]'; exercise benefits were discussed [technique] to target perceived benefits [construct]).
7.7b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.7c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.8a At least one, but not all, of the intervention techniques are explicitly linked to at least one theory-relevant construct/ predictor	Yes/No/Don't know At least one, but not all, of the intervention techniques (T) are explicitly linked to at least one theory-relevant construct/predictor (C). At least one technique, therefore, is not linked to any theory-relevant construct/predictor. Examples of meeting this criterion: - T1 and T2 are used to change C1; T3 is used but not linked to a construct/predictor (see DIAGRAM 2). - T1 used to change C1; T2 used to change C1 and C2; T3 is used but not linked to a construct/predictor. **DIAGRAM 2: Example: At least one, but not all, of the intervention techniques (T) are linked to at least one theory-relevant construct/predictor (C)**  In some cases, items 7, 8 and 9 may all be coded as 'no' (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa). If item 7 is coded 'yes' then items 8 and 9 must be coded 'no'. It is possible for items 8 and 9 to both be coded

	as ‘yes’.
7.8b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.8c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.9a Group of techniques are linked to a group of constructs/ predictors	Yes/No/Don’t know A cluster/group of techniques (T) is linked to a cluster/group of theory-relevant constructs/ predictors (C). Example of meeting this criterion: - T1, T2, T3 are used to change C1, C2 (see DIAGRAM 3). **DIAGRAM 3: Example: Group of techniques (T) are linked to a group of theory-relevant constructs/predictors (C)**  In some cases, items 7, 8 and 9 may all be coded as ‘no’ (e.g., where techniques are listed but theory-relevant constructs/predictors are not, or vice-versa). If item 7 is coded ‘yes’ then items 8 and 9 must be coded ‘no’. It is possible for items 8 and 9 to both be coded as ‘yes’.
7.9b List clusters of techniques/constructs and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.9c List clusters of techniques/predictors and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.10a All theory-relevant constructs/predictors are explicitly linked to at least one intervention	Yes/No/Don’t know Every theoretical construct within a stated theory, or every stated predictor (C) (see item 5), is linked to at least one intervention technique (T) Examples

technique	of meeting this criterion: - C1 & C2 are linked to T1; C3 is linked to T2. Constructs 1-3 are the only constructs within the theory specified in item 5 (see DIAGRAM 4). - C1 is linked to T1; C2 is linked to T2; C3 is linked to T3 To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the '*Table of Theories*'. **DIAGRAM 4: Example: All constructs within a stated theory/all predictors (C) are linked to at least one intervention technique (T)** 
7.10b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.10c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.11a At least one, but not all, of the theory relevant constructs/predictors are explicitly linked to at least one intervention technique	Yes/No/Don't know At least one, but not all, of the theoretical constructs within a stated theory or at least one, but not all, of the stated predictors (see item 5) are linked to at least one intervention technique.[5] At least one, but not all, of the constructs within a stated theory/all stated predictors (C) (see item 5) are linked to at least one intervention technique (T). At least one construct within a stated theory, therefore, is not linked to any intervention technique. Examples of meeting this criterion: - C1 and C2 are linked to T1; C3 is not linked to an intervention technique- or is not highlighted by the authors (see DIAGRAM 5).

	- C1 is linked to T1; C2 is linked to T1 and T2; C3 is not linked to an intervention technique- or is not highlighted by the authors. To determine which theoretical constructs should be identified and then linked to at least one intervention technique, please consult the '*Table of Theories*'. **DIAGRAM 5: Example: At least one, but not all, of the constructs within a stated theory/stated predictors (C) are linked to at least one intervention technique (T)** 
7.11b Construct (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.11c Predictor (list links) and location	List the supporting evidence and its location, as follows: ([insert page number], [insert paragraph number] or place reported in app).
7.12 Additional notes	If you have any other concerns/thoughts about this section, put them here.

C: QA_ PA BCTs: Behaviour Change Techniques with an evidence base for promoting/inhibiting physical activity promotion

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

PA BCTs extracted for item 3.15 will also be automatically copied across from data extraction tab to ease population of this spreadsheet.

8. PA BCTS

Using the data entered in QA_PA BCTs in the Data extraction tab, complete as follows:

- 8.1 Where a ‘Good BCT for PA’ has been used, place a 1 in the corresponding cell, UNLESS it’s one of the two following BCTs, in which case, put a 2 in the corresponding cell:
  - Behaviour practice/rehearsal
  - Prompts/cues
- 8.2 Where a ‘BCT with mixed evidence for PA’ has been used, place a 1 in the corresponding cell
- 8.3 Where a ‘Bad BCT for PA’ has been used, place a -1 in the corresponding cell

Totals for each set of BCTs will be automatically calculated. Do not enter any data into BCT cells that are not applicable as this will make the automatic calculation stop working.

D: QA_MARS: The Mobile App Rating Scale

It is important that you have downloaded the app, if available, at this point!

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

4. **App admin**

Item	**Explanation**
9.1 Version (sourced)	Specify the version that was tested – app details
9.2 Ratings	Provide the average star rating for the app that has been tested - both the App Annie ranking rating AND the app store star rating IF DIFFERENT
9.3 Number of ratings	Number of user ratings provided at the time of testing. For example, in Google Play, this number will be in the top right corner of the app page next to a person icon and the average star rating. Both the App Annie ratings AND the app store ratings IF DIFFERENT
9.4 Developer	Provide details (name) of the developers of the app – this might be in the paper, or, if the app is in the app store, it might be reported there (e.g. in Google Play this is at the bottom of the app page) or in the details of the app itself, e.g. try going to the ‘About’ menu item in the app.
9.5 Relevant expertise/credentials of development team (NON-MARS ITEM)	Provide any information available on the credentials or relevant expertise of the development team such as: • Behaviour change

	• Medicine • App development/programming • Physical activity/exercise
9.6 QA_Release date (NON-MARS ITEM)	Provide the release date for the app (reported in App Annie ranking document) otherwise enter NA
9.7 Last update	Capture the date it was last updated, even if this is BEYOND the version that is being tested. For example, in Google Play, this is at the bottom of the app page.
9.8 Cost	Capture the cost or if it was free to download.
9.9 Platform tested	Report the operating system (platform) that the app is being tested in. E.g. Android or iOS (Apple)
9.10 Affiliations	Enter all that apply: • Unknown • Commercial • Government • NGO • University Provide evidence for this and/or specify where this is reported.
9.11 Technical aspects	Enter all that apply: • Allows sharing (Facebook, Twitter, etc.) - specify • Has an app community (sharing with other users only) • Allows password-protection • Requires login • Sends reminders • Needs web access to function
9.12 Additional notes	If you have any other concerns/thoughts about this section, put them here.

FOR THE FOLLOWING SECTIONS, BASE YOUR ANSWERS ON THE FINAL VERSION OF THE APP that has been downloaded. Do not use developmental information to complete this section unless it referring to the final iteration of the app described in the paper, i.e. no further changes were made to the app following feedback/comments/piloting of the app.

The Rating scale assesses app quality on four dimensions. All items are rated on a 5-point scale from “1.Inadequate” to “5.Excellent” but some items allow you to answer NA too. Enter the number that most accurately represents the quality of the app component you are rating. Please use the descriptors provided for each response category.

5. Engagement

Engagement – fun, interesting, customisable, interactive (e.g. sends alerts, messages, reminders, feedback, enables sharing), well-targeted to audience

Item	**Explanation**
10.1a Entertainment: Is the app fun/entertaining to use? Does it use any strategies to increase engagement through entertainment (e.g. through gamification)? Explain answer	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Dull, not fun or entertaining at all 2. Mostly boring 3. OK, fun enough to entertain user for a brief time (< 5 minutes) 4. Moderately fun and entertaining, would entertain user for some time (5-10 minutes total) 5. Highly entertaining and fun, would stimulate repeat use Provide evidence for your coding.
10.1b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
10.2a Interest: Is the app interesting to use? Does it use any strategies to increase engagement by presenting its content in an interesting way?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Not interesting at all 2. Mostly uninteresting 3. OK, neither interesting nor uninteresting; would engage user for a brief time (< 5 minutes) 4. Moderately interesting; would engage user for some time (5-10 minutes total) 5. Very interesting, would engage user in repeat use

	Provide evidence for your coding.
10.2b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
10.3a Customisation: Does it provide/retain all necessary settings/preferences for apps features (e.g. sound, content, notifications, etc.)?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Does not allow any customisation or requires setting to be input every time 2. Allows insufficient customisation limiting functions 3. Allows basic customisation to function adequately 4. Allows numerous options for customisation 5. Allows complete tailoring to the individual's characteristics/preferences, retains all settings Provide evidence for your coding.
10.3b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.

10.4a Interactivity: Does it allow user input, provide feedback, contain prompts (reminders, sharing options, notifications, etc.)? Note: these functions need to be customisable and not overwhelming in order to be perfect.	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No interactive features and/or no response to user interaction 2. Insufficient interactivity, or feedback, or user input options, limiting functions 3. Basic interactive features to function adequately 4. Offers a variety of interactive features/feedback/user input options 5. Very high level of responsiveness through interactive features/feedback/user input options Provide evidence for your coding.
10.4b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
10.5a Target group: Is the app content (visual information, language, design) appropriate for your target audience?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Completely inappropriate/unclear/confusing 2. Mostly inappropriate/unclear/confusing 3. Acceptable but not targeted. May be inappropriate/unclear/confusing 4. Well-targeted, with negligible issues 5. Perfectly targeted, no issues found Provide evidence for your coding.
10.5b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted?

	Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
10.6 Engagement mean score	Do not enter anything here

6. Functionality

Functionality – app functioning, easy to learn, navigation, flow logic, and gestural design of app

Item	**Explanation**
11.1 Performance: How accurately/fast do the app features (functions) and components (buttons/menus) work?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. App is broken; no/insufficient/inaccurate response (e.g. crashes/bugs/broken features, etc.) 2. Some functions work, but lagging or contains major technical problems 3. App works overall. Some technical problems need fixing/Slow at times 4. Mostly functional with minor/negligible problems 5. Perfect/timely response; no technical bugs found/contains a 'loading time left' indicator Provide evidence for your coding.
11.2 Ease of use: How easy is it to learn how to use the app; how clear are the menu labels/icons and instructions?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No/limited instructions; menu labels/icons are confusing; complicated 2. Useable after a lot of time/effort 3. Useable after some time/effort 4. Easy to learn how to use the app (or has clear instructions) 5. Able to use app immediately; intuitive; simple Provide evidence for your coding.

11.3 Navigation: Is moving between screens logical/accurate/appropriate/ uninterrupted; are all necessary screen links present?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Different sections within the app seem logically disconnected and random/confusing/navigation is difficult 2. Usable after a lot of time/effort 3. Usable after some time/effort 4. Easy to use or missing a negligible link 5. Perfectly logical, easy, clear and intuitive screen flow throughout, or offers shortcuts Provide evidence for your coding.
11.4 Gestural design: Are interactions (taps/swipes/pinches/scrolls) consistent and intuitive across all components/screens?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Completely inconsistent/confusing 2. Often inconsistent/confusing 3. OK with some inconsistencies/confusing elements 4. Mostly consistent/intuitive with negligible problems 5. Perfectly consistent and intuitive Provide evidence for your coding.
11.5 Functionality mean score	Do not enter anything here

12. Aesthetics

Aesthetics – graphic design, overall visual appeal, colour scheme, and stylistic consistency

Item	Explanation
12.1 Layout: Is arrangement and size of buttons/icons/menus/content on the screen appropriate or zoomable if needed?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Very bad design, cluttered, some options impossible to select/locate/see/read. Device display not optimised 2. Bad design, random, unclear, some options difficult to select/locate/see/read 3. Satisfactory, few problems with selecting/locating/seeing/reading items or with minor screen-size problems 4. Mostly clear, able to select/locate/see/read items 5. Professional, simple, clear, orderly, logically organised, device display optimised. Every

	design component has a purpose Provide evidence for your coding.
12.2 Graphics: How high is the quality/resolution of graphics used for buttons/icons/menus/content?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Graphics appear amateur, very poor visual design - disproportionate, completely stylistically inconsistent 2. Low quality/low resolution graphics; low quality visual design – disproportionate, stylistically inconsistent 3. Moderate quality graphics and visual design (generally consistent in style) 4. High quality/resolution graphics and visual design – mostly proportionate, stylistically consistent 5. Very high quality/resolution graphics and visual design - proportionate, stylistically consistent throughout Provide evidence for your coding.
12.3 Visual appeal: How good does the app look?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. No visual appeal, unpleasant to look at, poorly designed, clashing/mismatched colours 2. Little visual appeal – poorly designed, bad use of colour, visually boring 3. Some visual appeal – average, neither pleasant, nor unpleasant 4. High level of visual appeal – seamless graphics – consistent and professionally designed 5. As above + very attractive, memorable, stands out; use of colour enhances app features/menus Provide evidence for your coding.
12.4 Aesthetics mean score	Do not enter anything here

13. Information

Information – Contains high quality information (e.g. text, feedback, measures, and references) from a credible source. Select N/A if the app component is irrelevant.

Item	Explanation
13.1 Accuracy of app description (in app store): Does app contain what is described?	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Misleading. App does not contain the described components/functions. Or has no description 2. Inaccurate. App contains very few of the described components/functions 3. OK. App contains some of the described components/functions 4. Accurate. App contains most of the described components/functions 5. Highly accurate description of the app components/functions Provide evidence for your coding.
13.2a Goals: Does app have specific, measurable and achievable goals (specified in app store description or within the app itself)?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A Description does not list goals, or app goals are irrelevant to research goal (e.g. using a game for educational purposes) 1. App has no chance of achieving its stated goals 2. Description lists some goals, but app has very little chance of achieving them 3. OK. App has clear goals, which may be achievable. 4. App has clearly specified goals, which are measurable and achievable 5. App has specific and measurable goals, which are highly likely to be achieved Provide evidence for your coding.
13.2b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour

	Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
13.3a Quality of information: Is app content correct, well written, and relevant to the goal/topic of the app?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A There is no information within the app 1. Irrelevant/inappropriate/incoherent/incorrect 2. Poor. Barely relevant/appropriate/coherent/may be incorrect 3. Moderately relevant/appropriate/coherent/and appears correct 4. Relevant/appropriate/coherent/correct 5. Highly relevant, appropriate, coherent, and correct Provide evidence for your coding.
13.3b What promoted behaviour does your previous answer relate to?	Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted? Enter ONE of the below: • Physical activity only • Other promoted behaviour(s) NOT including physical activity • Both physical activity AND at least one other promoted behaviour Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.
13.4a Quantity of information: Is the extent coverage within the scope of the app; and comprehensive but concise?	Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A There is no information within the app 1. Minimal or overwhelming 2. Insufficient or possibly overwhelming 3. OK but not comprehensive or concise 4. Offers a broad range of information, has some gaps or unnecessary detail; or has no links to more information and resources 5. Comprehensive and concise; contains links to more information and resources

<table>
<tr><td></td><td>Provide evidence for your coding.</td></tr>
<tr><td>13.4b What promoted behaviour does your previous answer relate to?</td><td>Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted?

Enter ONE of the below:
Physical activity only
Other promoted behaviour(s) NOT including physical activity
Both physical activity AND at least one other promoted behaviour
Where other promoted behaviours include but are not limited to things like a healthy diet, smoking, etc.</td></tr>
<tr><td>13.5a Visual information: Is visual explanation of concepts – through charts/graphs/images/videos, etc. – clear, logical, correct?</td><td>Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app:

N/A There is no visual information within the app (e.g. it only contains audio, or text)
Completely unclear/confusing/wrong or necessary but missing
Mostly unclear/confusing/wrong
OK but often unclear/confusing/wrong
Mostly clear/logical/correct with negligible issues
Perfectly clear/logical/correct
Provide evidence for your coding.</td></tr>
<tr><td>13.5b What promoted behaviour does your previous answer relate to?</td><td>Is your previous answer based on strategies/techniques that relate to a specific behaviour or multiple behaviours e.g. were fun techniques used to promote a healthy diet rather than physical activity promotion, or were they used for all behaviours that were being promoted?

Enter ONE of the below:
Physical activity only
Other promoted behaviour(s) NOT including physical activity
Both physical activity AND at least one other promoted behaviour
Where other promoted behaviours include but are not limited to things like a healthy diet, smoking,</td></tr>
</table>

	etc.
13.6 Credibility: Does the app come from a legitimate source (specified in app store description or within the app itself)?	For this item, you may need to go online to check the legitimacy of the source. Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Source identified but legitimacy/trustworthiness of source is questionable (e.g. commercial business with vested interest) 2. Appears to come from a legitimate source, but it cannot be verified (e.g. has no webpage) 3. Developed by small NGO/institution (hospital/centre, etc.) /specialised commercial business, funding body 4. Developed by government, university or as above but larger in scale 5. Developed using nationally competitive government or research funding (e.g. Australian Research Council, NHMRC) Provide evidence for your coding.
13.7 Evidence base: Has the app been trialled/tested; must be verified by evidence (in published scientific literature)?	For this item, you may have been provided with additional papers, if not, base your answer on the information provided in the original paper. Enter ONE of the numbers, or NA, below that best corresponds to your perception of the app: N/A The app has not been trialled/tested 1. The evidence suggests the app does not work 2. App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has partially positive outcomes in studies that are not randomised controlled trials (RCTs), or there is little or no contradictory evidence. 3. App has been trialled (e.g., acceptability, usability, satisfaction ratings) and has positive outcomes in studies that are not RCTs, and there is no contradictory evidence. 4. App has been trialled and outcome tested in 1-2 RCTs indicating positive results 5. App has been trialled and outcome tested in > 3 high quality RCTs indicating positive

	results Provide evidence for your coding.
13.8 Information mean score	Do not enter anything here.

14. App quality mean score

This will be automatically calculated, do not enter anything here.

(NOTE TO LL: * Exclude questions rated as “N/A” from the mean score calculation.)

15. Perceived Impact

These added items can be adjusted and used to assess the perceived impact of the app on the user’s knowledge, attitudes, and intentions to change as well as the likelihood of actual change in the target health behaviour.

Item	Explanation
15.1 Awareness: This app is likely to increase awareness of the importance of addressing physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree Provide evidence for your coding.
15.2 Knowledge: This app is likely to increase knowledge/understanding of physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree Provide evidence for your coding.
15.3 Attitudes: This app is likely to change attitudes toward improving physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3.

	4. 5. Strongly agree Provide evidence for your coding.
15.4 Intention to change: This app is likely to increase intentions/motivation to address physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree Provide evidence for your coding.
15.5 Help seeking: Use of this app is likely to encourage further help seeking for physical activity (if it's required)	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree Provide evidence for your coding.
15.6 Behaviour change: Use of this app is likely increase/decrease physical activity	Enter ONE of the numbers below that best corresponds to your perception of the app: 1. Strongly disagree 2. 3. 4. 5. Strongly agree Provide evidence for your coding.

E: QA_PA measure reliability: Reliability and validity of the measures of physical activity

Admin details

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

3.6/3.7 PA measure

WILL BE AUTOMATICALLY COPIED FROM DATA EXTRACTION TAB

16. Evaluation of PA measure

Item	Explanation
16.1 Evidence for/against measure?	If paper/website/app reports any evidence for/against the measure, report it here. This includes any measures of reliability or validity that may be reported.
16.2 Strengths of measure	Provide details of any strengths of the measure reported in the paper/website/app, or any you perceive. For example any preferences users had for the measure, ease of use, brevity, perceived accuracy etc. If it's a self-report measure ALSO enter: 'self-report measure strengths' If it's an objective measure ALSO enter: 'objective measure strengths'
16.3 Weaknesses of measure	Provide details of any weaknesses of the measure reported in the paper/website/app, or any you perceive. For example any difficulties the users had with the measure, functional problems or misunderstandings or lack of wear/completion etc. If it's a self-report measure ALSO enter: 'self-report measure weaknesses' If it's an objective measure ALSO enter: 'objective measure weaknesses'
16.4 Additional notes	If you have any other concerns/thoughts about this section, put them here.

F: QA_50 most recent user reviews

These will have been extracted by LL and do not need processing by second coder

Appendices

NB. Appendices 1, 2, 3 and 6 are not attached as they refer to separate book chapters/books/papers that can be accessed

Appendix 1 – Taylor, N. 2014. Challenges in measuring physical activity in the context of mental health *In*: A. Clow and S. Edmunds, eds. *Physical Activity and Mental Health*. Leeds: Human Kinetics, pp.41–61.

Appendix 2 - Strath, S.J., Kaminsky, L.A., Ainsworth, B.E., Ekelund, U., Freedson, P.S., Gary, R.A., Richardson, C.R., Smith, D.T. and Swartz, A.M. 2013. Guide to the assessment of physical activity: Clinical and research applications: A scientific statement from the American Heart association. *Circulation*. **128**(20), pp.2259–2279.

Appendix 3 – Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M.P., Cane, J. and Wood, C.E. 2013. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine*. [Online]. **46**(1), pp.81–95. Available from: http://link.springer.com/10.1007/s12160-013-9486-6.

Appendix 4 – PA BCTs (below)

Appendix 5 – Quality indicators and definitions (below)

Appendix 6 - Michie, S., West, R., Campbell, R., Brown, J. and Gainforth, H. 2014. *ABC of behaviour change theories: an essential resource for researchers, policy makers and practitioners* [Online]. Silverback Publishing. Available from: https://www.waterstones.com/book/abc-of-behaviour-change-theories/susan-michie/robert-west/9781912141012.

Appendix 4 – BCTs associated with changes in physical activity

Positive	Negative	Mixed evidence
2.6 Biofeedback	4.2 Information about antecedents	1.2 Problem solving
6.1 Demonstration of behaviour		1.5 Reviewing behavioural goal
8.1 Behaviour Practice/rehearsal		2.2 Feedback on behaviour
8.7 Graded tasks		
1.4 Action planning		
4.1 Instruction on how to perform behaviour		
7.1 Prompts/cues		
10.9 Self-reward		
2.3 Self-monitoring of behaviour		
5.1 Providing info on health consequences		
8.2 Behaviour substitutions		
12.1 Restructuring physical environment		
3.2 Providing practical social support		
12.2 Restructuring social environment		
3.1 Unspecified social support		
1.3 Setting outcome goals		
9.2 Pros and cons		
10.4 Social rewards		
8.3 Habit formation		
1.9 Commitment		
1.6 Discrepancy between current behaviour and goals		

Appendix 5 – Quality indicators and definitions

	Quality indicator	Summary definition
1	Acceptability or Participant perceptions	Positive and negative feedback or recommendations from users on the app content e.g. preferences or recommendations for information, tone or features. Can include user ratings or reviews of the app (the latter may overlap with other indicators such as 3, or 4, these should be coded separately). May include barriers and facilitators (feasibility) to use of the app and/or smartphone such as it being easy to fit self-report requests into your routine, or forgetting to carry the phone. May also include whether or not users/providers/practitioners would recommend the app to others/patients
2	Aesthetics	Visual attractiveness of the app interface design in terms of colours, fonts, and layout. How professional the design is. How pleasing to eye the design and layout is. Can include relevance of design to the behaviour.
3	Credible, trustworthy/appropriate or useful/essential information	Content of the app is likely to be accurate or believable – not making impossible or implausible claims. Content is safe for users, won't harm them or will minimise harm or provides a caveat for medical information that requires seeing a professional. Information/app appears useful.
4	Engagement	Use of methods to encourage user interactivity with the app, can include use of certain strategies or features that promote/inhibit for example, feedback, tailoring, prompts/reminders, gamification. (Often this term has also been used to refer to usage/response to app intervention features or feasibility, such as required step-count submissions or required message responses (Monroe et al., 2015). Or it has encompassed both these and the above summary definition (Rose et al., 2017). These items have been separated out here).
5	Functionality/Usability	Ease of use of the app and/or smartphone features, such as navigation, terminology, design in relation to ease of use, not aesthetics (see 4) as well as general perception of how much support might be required for use or how complex or inconsistent it might be. Functional errors related to app operations such as

		bugs/crashing also captured here. Includes practicality of use for promoting or capturing physical activity based on functions and features. Can be assessed by questionnaires such as the System Usability Scale, interviews or user-testing/performance tests. In one instance, presence of the app in the top 100 of a category of the app store was also suggested as a potential proxy for usability.

Appendix 16 Targeted behaviours of apps

App name	Targeted behaviour (PA only or PA+)	Type of physical activity	Other behaviours targeted/captured
Haptivity	PA	User defined - any activity	No
MAPS	PA	Moderate-to-vigorous physical activity (MVPA), minutes per day. User specified, but aerobic and non-aerobic specified in Goal setting condition.	No
Unnamed app	PA+	Not explicitly specified, but appears to be anything where user will be travelling.	Sleep, socialising and mood
ATHENA	PA+	Exercise and physical activity and steps (mentions basketball, tennis, badminton captured by wearable, outdoor activities by activity recognition such as walking, jogging, riding on bus/subway, staying still).	Captures sleep, mental health, mood, social wellbeing. Also suggests that recommendations include food.
Health Mashups	PA+	Pilot - step count and amount of exercise. Trial - step count only.	Weight (daily food intake, weight and body fat - latter reported in pilot paper), sleep (hours slept, time awoken), mood and pain. Contextual information that may impact wellbeing - weather, location (city level), busyness (hours busy per day).
iN Touch	PA+	User-specified	Diet, mood, socialising
Motimate	PA+	User defined	Weight loss, food, mood, stress
Ngala	PA+	Not reported	Weight, diet, emotional wellbeing and

			sleep patterns.
SIGMA	PA+	Steps per day	Weight maintenance/weight loss - (1) self-reported maladaptive thoughts related to eating and body weight, (2) self-reported maladaptive eating behaviours in the range of urgent food cravings, emotional eating or binge eating, (3) as well as biased attentional processing of food items as indexed by reaction times. Secondary outcomes will be represented by changes in weight, Body Mass Index, general mood.
Descriptives:	**PA focus: 2/9** **PA+: 7/9**		
One You	PA	Running	No
Runkeeper	PA	User-specified but only from app list which includes but is not limited to walking and other, or from free training plans which are running-based.	Also tracks weight.
Runtastic HR	PA+	Not specified - just whether or not user is/has been/about to be active, e.g. resting, general, pre/post sport, max HR	Additional behaviours discussed at whim on website and through News item on app e.g. going plastic free, nutrition.
Runtastic	PA+	Wide range of specified activity types from aerobics to Zumba, including walking. User can add generic 'other'. Can also specify workout goal for some activities where distance and duration goals are applicable.	Additional behaviours discussed at whim on website and through News item on app e.g. going plastic free, nutrition.
Runtastic PRO	PA+		
Fitbit	PA+	User specified activities and steps. Coach app also tracks app specific	Weight, water, diet.

		workouts.	
Keep	PA+	Steps and activities listed only and training plans only.	Calories/nutrition
Clue	PA+	Exercise	Tracks periods, craving, digestion, fluid, hair, pain, skin, stool, temperature, weight, emotions, energy, mental, motivation, sleep, social, appointments, party, sex, sickness and contraceptives. Each has an information page about it explaining why it's relevant to the monthly cycle and what is healthy to do/experience, therefore considered to be promoting healthy behaviours.
Samsung Health	PA+	PA, exercise, active time, steps	Water intake, weight, calories, caffeine, sleep.
Up app	PA+	Steps	Sleep
Descriptives:	**PA focus: 2/10** **PA+: 8/10**		
TOTAL	**PA focus: 4/19** **PA+: 15/19**		

PA = physical activity only; PA+ = physical activity and other behaviours targeted

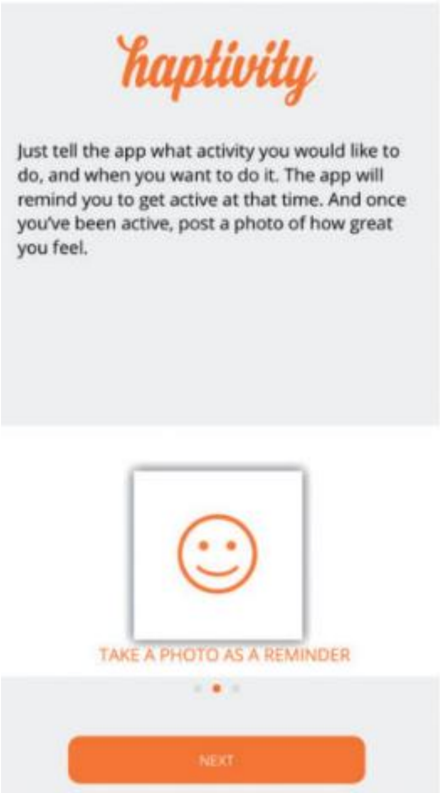
Appendix 17 Screenshots of methods of capture and feedback of affect

Screenshots/images of mood capture and feedback for apps, where available.

Literature-based apps

Haptivity (Forster et al., 2017)

A. Instructions for capturing affect



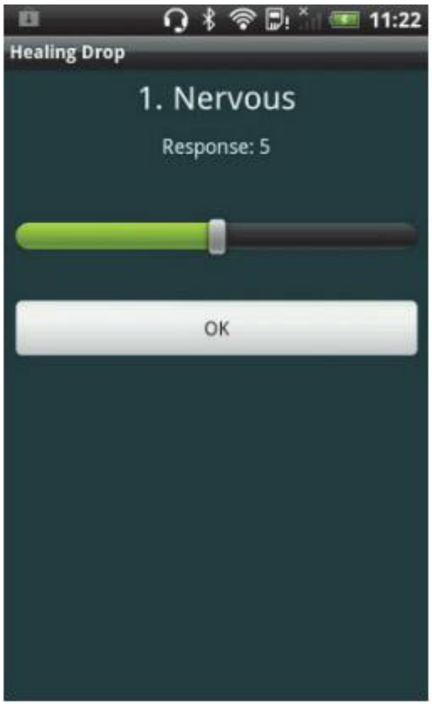
MAPS (Fanning et al., 2017)

A. Capturing affect

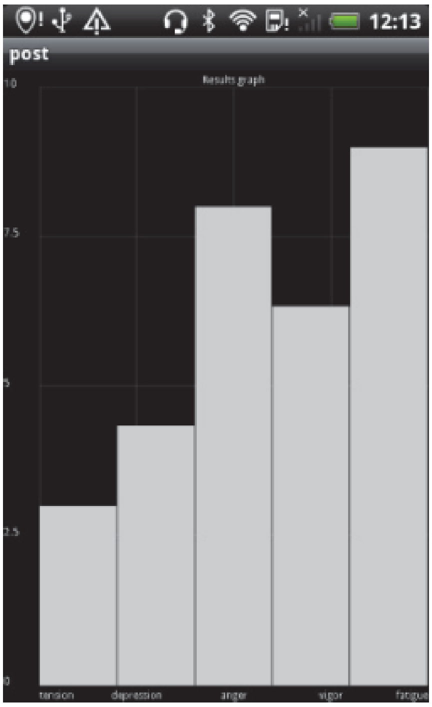
B. Feedback on affect

Unnamed (Fernández et al., 2013)

A. Capturing affect

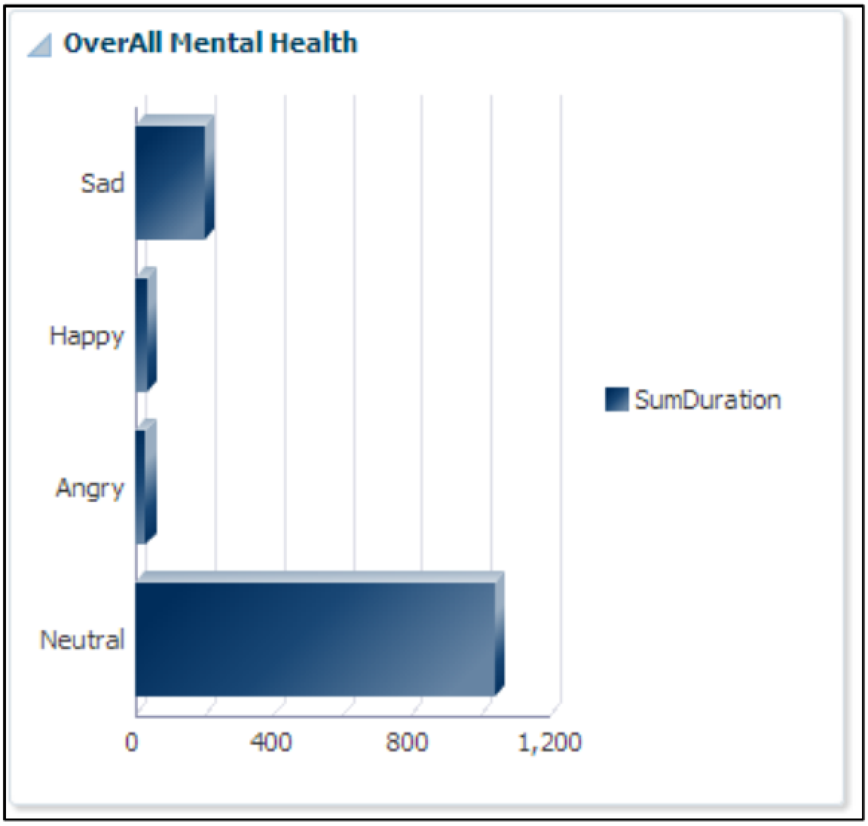


B. Feedback on affect

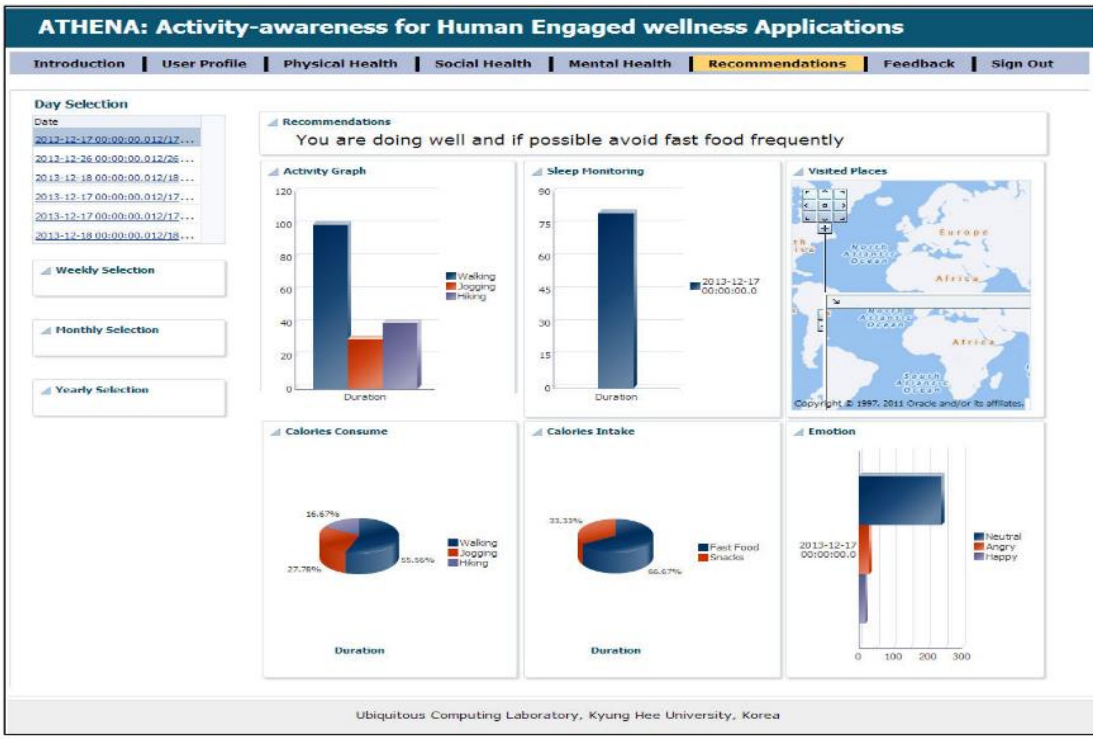


ATHENA (Fahim et al., 2014)

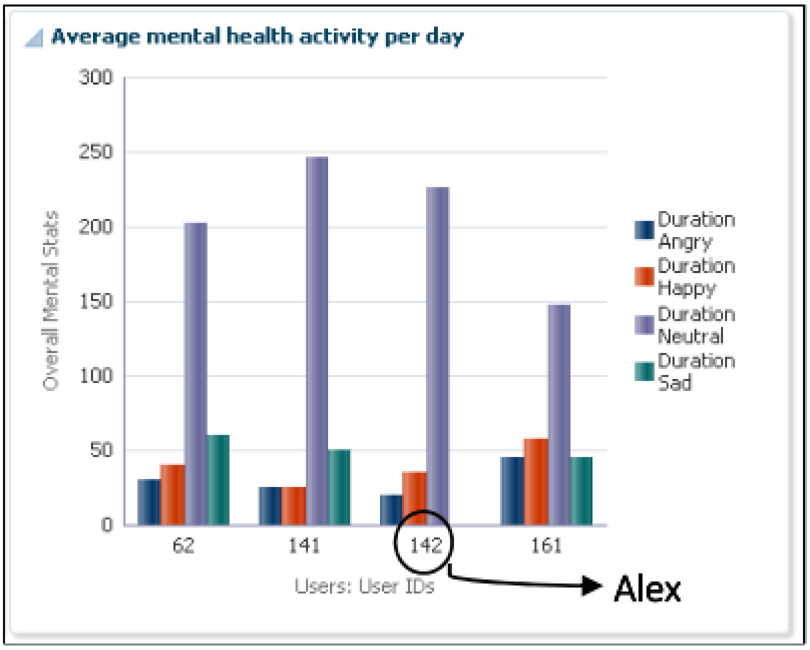
A. Feedback on affect



B. Feedback on affect



C. Feedback on affect compared to other users



Health Mashups (Bentley et al., 2013)

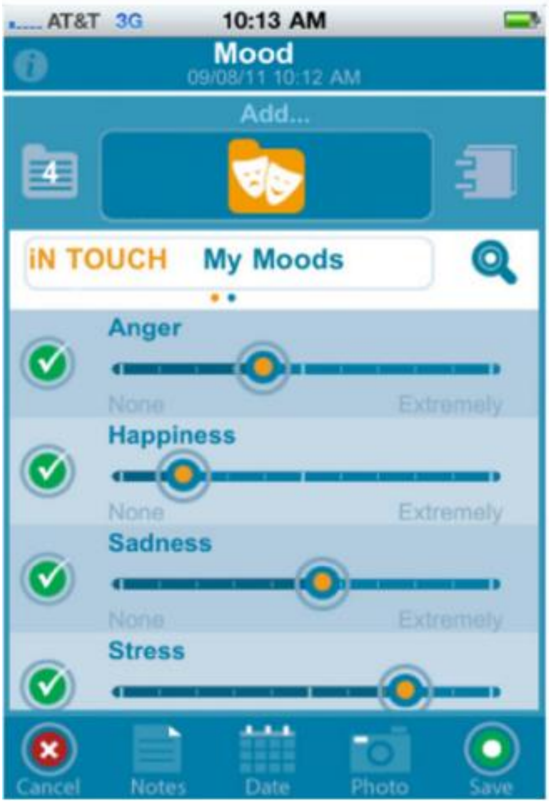
A. Feedback on affect

B. Feedback on affect

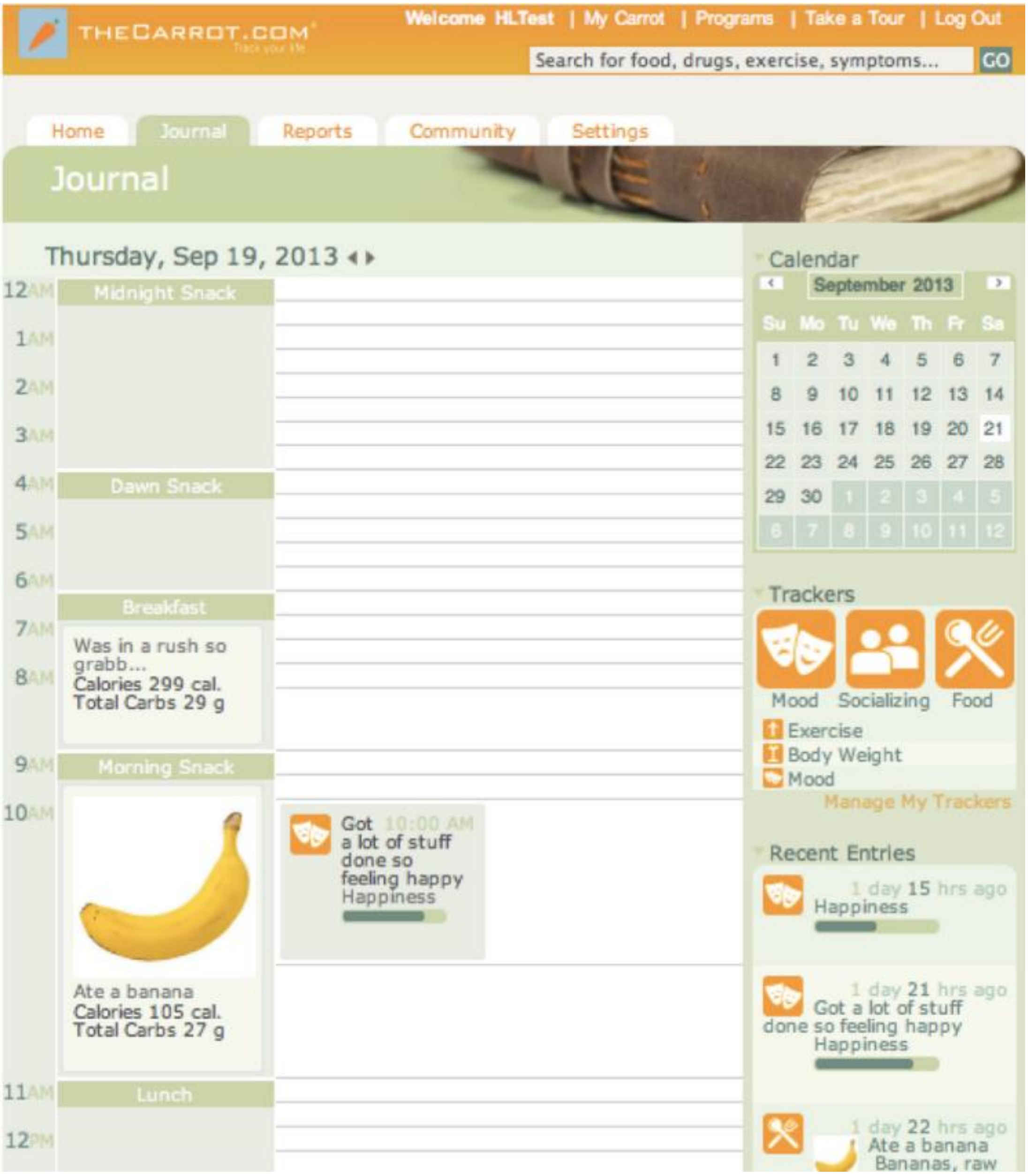
C. Feedback on affect

iN Touch (Kim et al., 2015)

A. Capturing affect

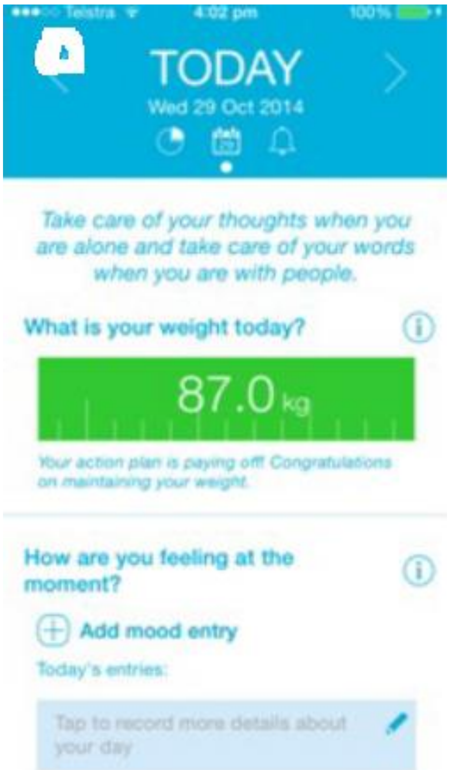


B. Feedback on affect

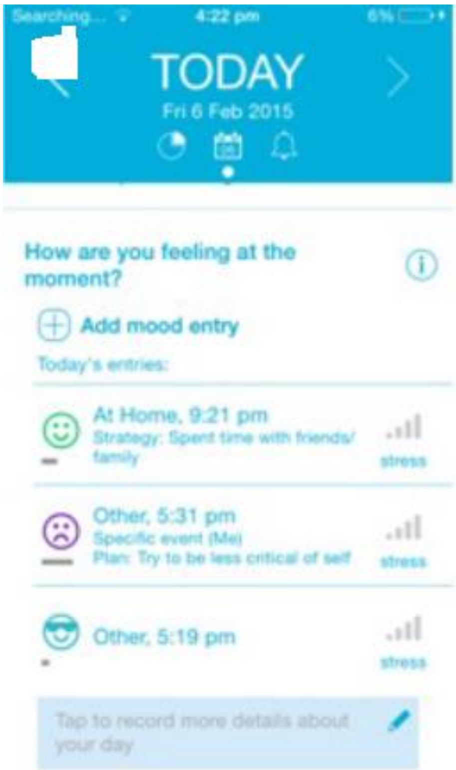


Motimate (Brindal et al., 2016)

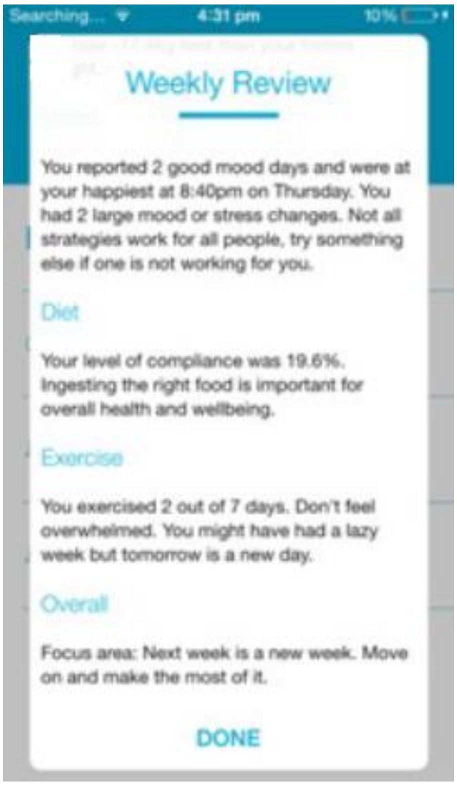
A. Capturing affect



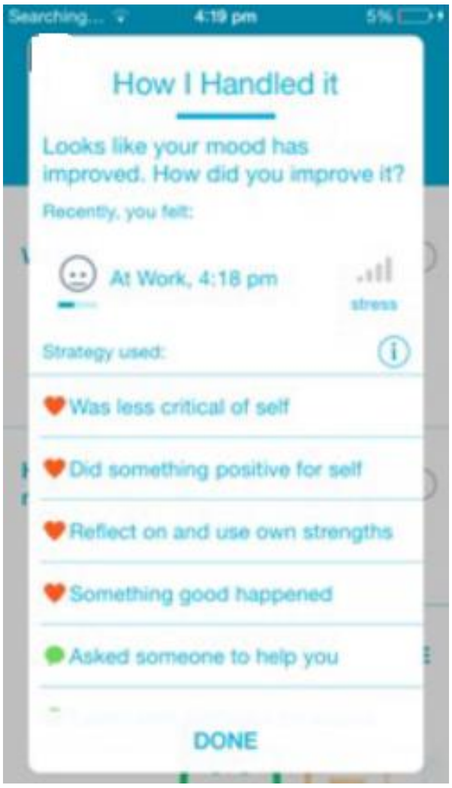
B. Capturing/Feedback on affect



C. Feedback on affect



D. Feedback on affect



Ngala (Hearn et al., 2014)

No images available

SIGMA (Podina et al., 2017)

No images available

Public apps

One You

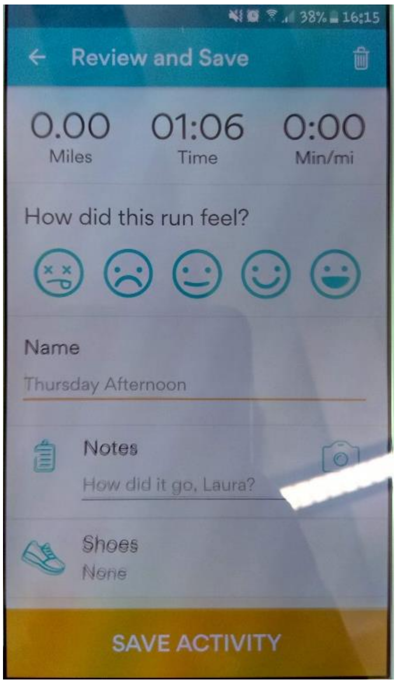
A. Capturing affect pre-run

B. Capturing affect post-run

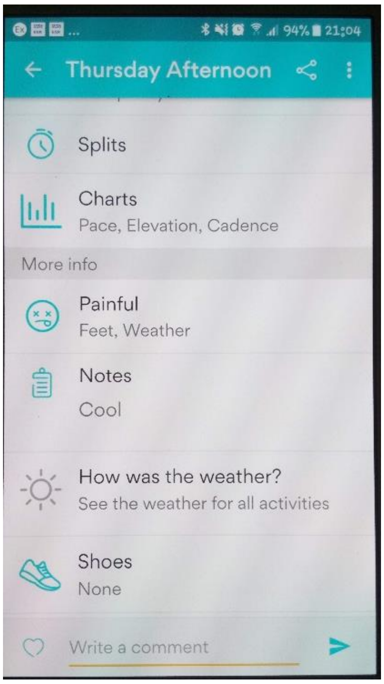
C. Feedback on affect

Runkeeper

A. Capturing affect post-run

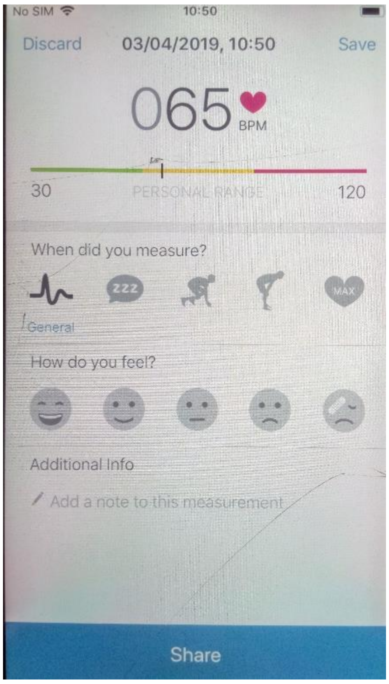


B. Feedback on affect

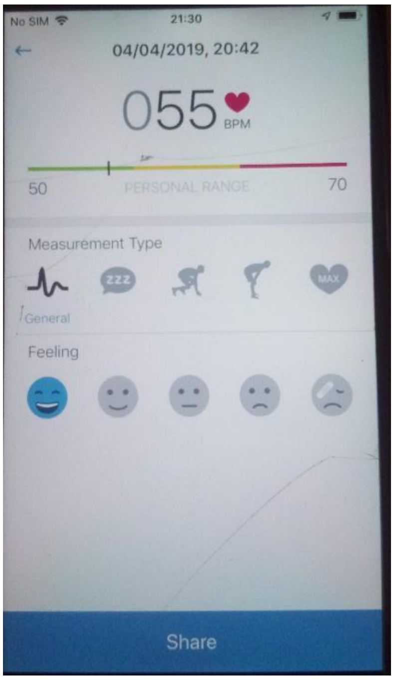


Runtastic HR

A. Capturing affect post-HR reading



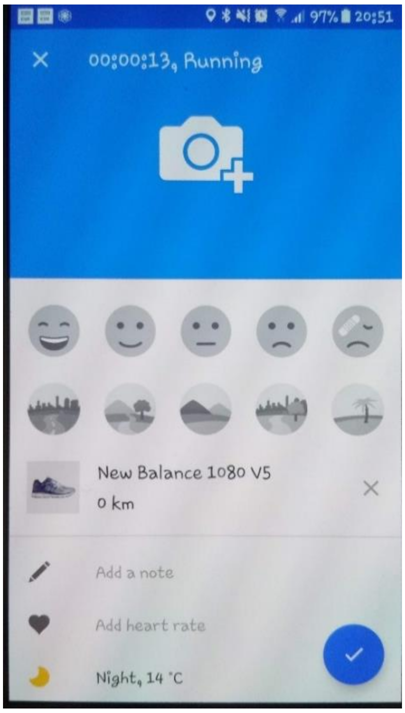
B. Feedback on affect

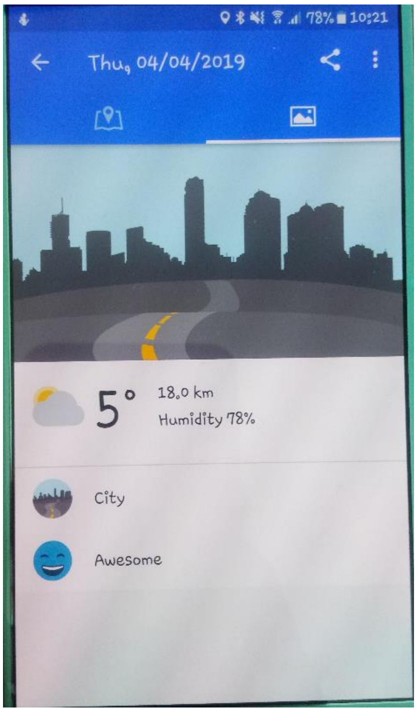


Runtastic and Runtastic PRO

A. Capturing affect post-run

B. Feedback on affect

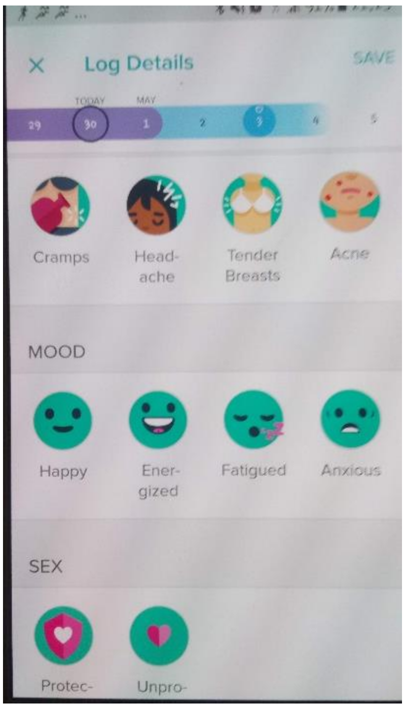


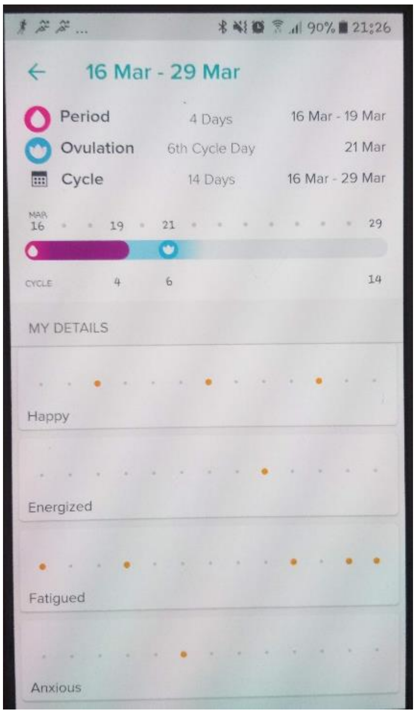


Fitbit

A. Capturing affect

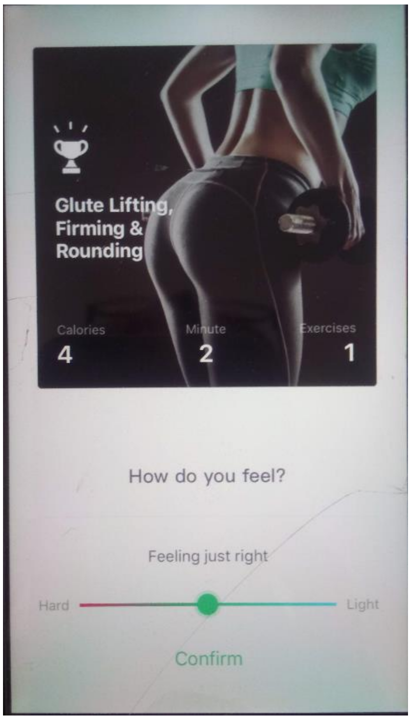
B. Feedback on affect



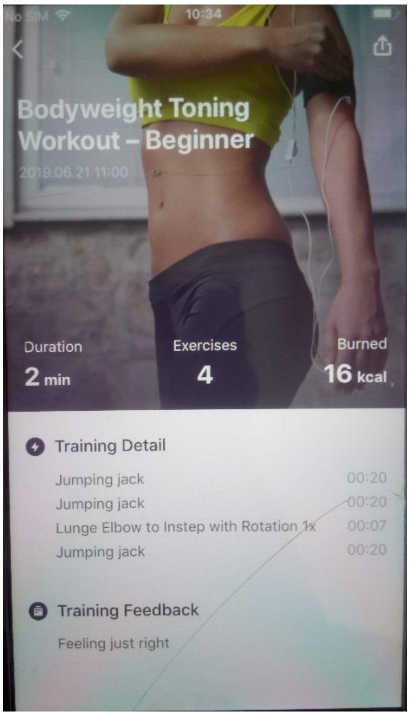


Keep

A. Capturing affect post-exercise

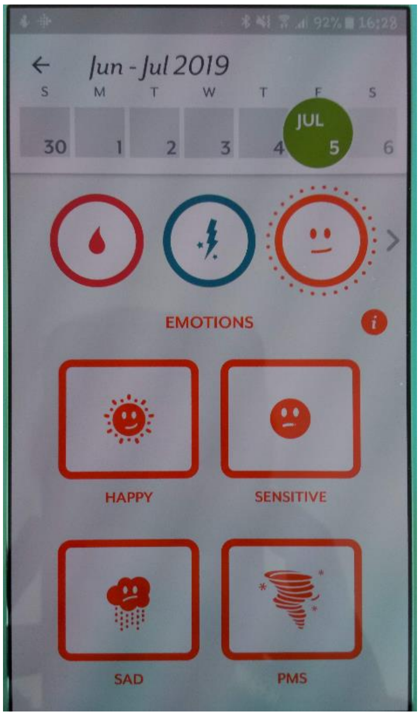


B. Feedback on affect

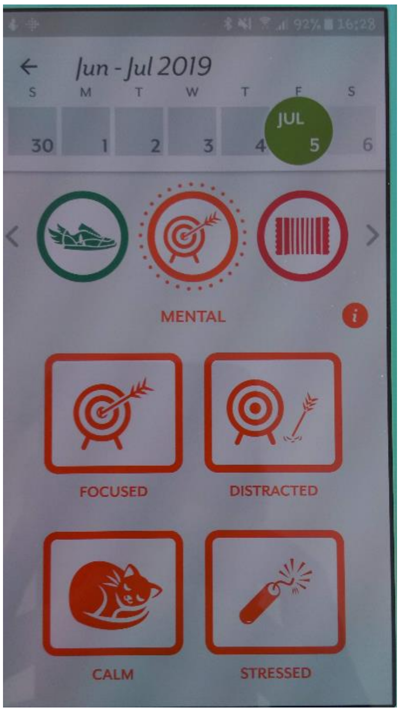


Clue

A. Capturing affect (Emotions)

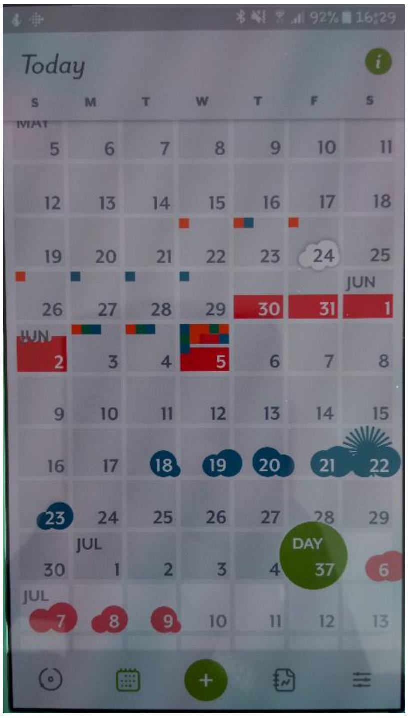


B. Capturing affect (Mental)

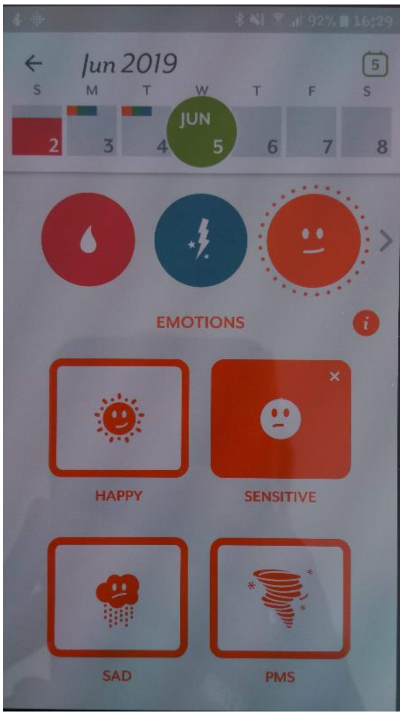


C. Feedback on affect (Emotions)

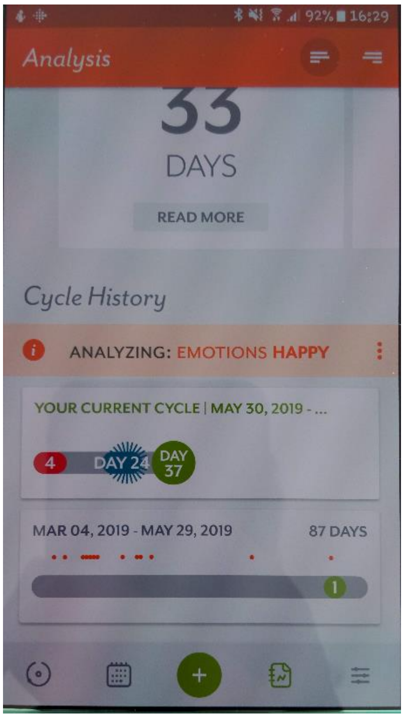
(Daily records indicated by orange square)



D. Feedback on affect

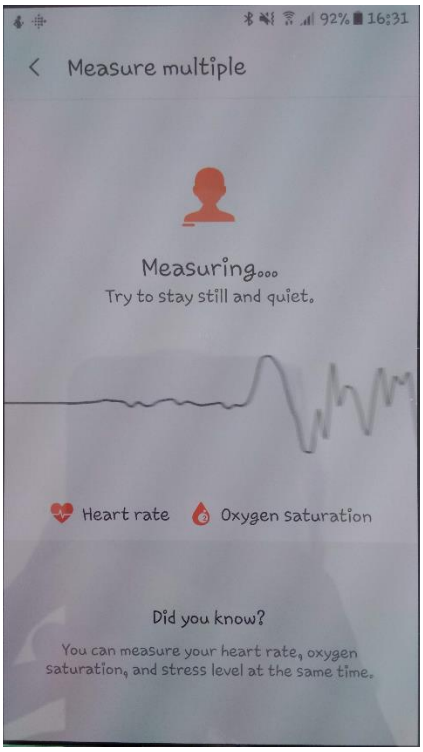


E. Feeback on affect (During menstrual cycle)

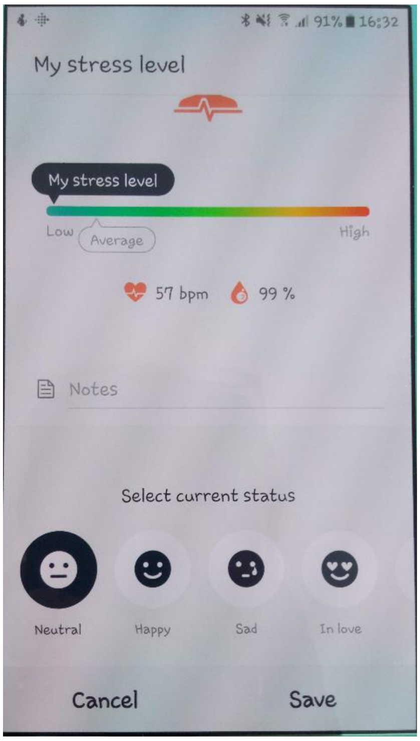


Samsung Health

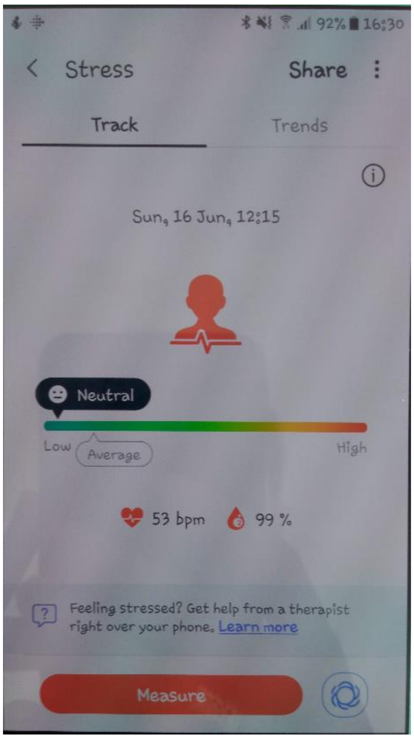
A. Capturing affect
(Stress level via heart rate)



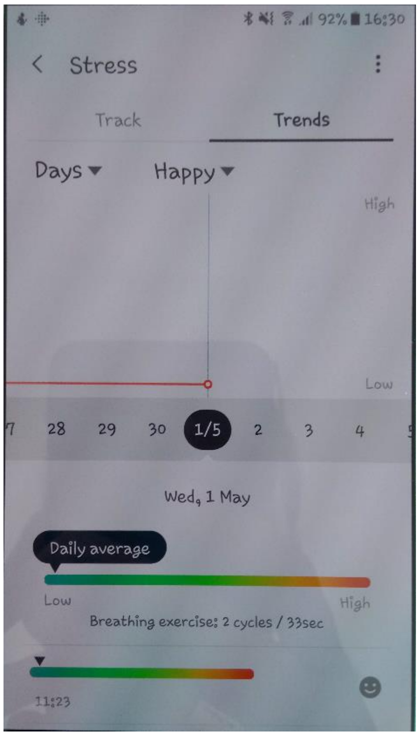
B. Capturing affect



C. Feedback on affect



D. Feedback on affect



Up (Lyons et al., 2014)

A. Capturing affect

Appendix 18 Extra evidence that informed MARS scores

Additional evidence of quality that informed MARS scores for the apps in addition to the app itself, or the app description and screenshots provided in papers.

Text from additional evidence about app quality occasionally features multiple times in the following tables. This is because it features information for multiple quality criteria. Extra text has been retained to maintain the context around the statements for better clarity.

App		Additional evidence for aesthetics	Additional evidence for credibility	Additional evidence for effectiveness and potential impact	Additional evidence for engagement	Additional evidence for functionality
Haptivity (Not available for download) PA		Assessed but not reported	The ability to quantify the PA performed, whilst not essential for some users, would enhance the app's value for others. (Forster et al., 2017)	The encouragement element of the app created motivation for some users. The reminder function appeared to be successful in encouraging activity.	NA	Assessed but not reported
MAPS (Not available for download)	1 2 3	NA	Users made positive comments about graphs being clear/informative (conditions 2/4) (see table 17).	NA	Knowledge videos perceived as too long (all conditions) (see table 17).	Among individuals with access to in-app goal setting (groups 2 and 4 (see table 17)), 58% reported that tracking goal-related activities

PA	4		The points group perceived the point's component to provide helpful feedback. Knowledge module seen as informative by all groups but users also felt that they knew all the information.(Fanning et al., 2017)			was very easy, and another 28% reported it was fairly easy. For non-goal activities, which required manual entry of an activity name, 38% reported it was very easy, 22% reported it was fairly easy, and 24% reported they felt neutral. For the in-app goal setting process, 50% reported it was very easy and 22% reported it was fairly easy. For individuals without access to in-app goal setting (groups 1 and 3(see table 17)), 51% reported that tracking goal-relate activities was very easy, while 17% reported it was fairly easy. For non-goal activities, 36% reported it was very easy to enter these activities, 15% reported it was fairly

						easy, 19% reported they were neutral, and another 19% felt it was a little challenging. For the goal-setting process via the goal setting handbook, 15% reported it was very easy, 32% reported it was fairly easy, 21% reported they felt neutral, and 28% indicated it was fairly challenging. Positive responses were provided with respect to the ease of use of tracking (across all groups), in-app goals (groups 2&4) and the knowledge videos were easy to watch (all groups). The following negative responses were reported: lack of retrospective tracking (all groups), inability to edit goals within the

					week (2&4), lack of in-depth goals, lack of frequent activities menu, and lack of in-app goals (1&3).
Unnamed app (Not available for download) PA+	NA	NA	NA	NA	NA
ATHENA apps (Not available for download) PA+	NA	The accuracy of correct recommendations for food, physical and mental therapies was 98.7% and was stable. Good results were generated by the hybrid reasoner (which generates recommendations) where the individual results from all the three learners are combined together, however individual accuracy's of each	NA	NA	NA

		learner varies with some outperforming others.(Ali et al., 2015)			
Health Mashups (Not available for download) PA+	NA	**Pilot outcomes** Participants found certain entries in their feeds to be interesting and they were able to learn more about themselves through the app and mobile website. The day of the week observations were also quite useful to better understand trends over time. Many individual correlations made sense to users. The “scientific” nature of the system also appealed to participants, with comments including that it was “very truthful” and they	NA	**Pilot outcomes** Repetitive observations discouraged frequent use because most of the same observations remained significant from day to day. Many participants reported wanting something “new.” Some participants suggested using alerts when new observations were available. Other participants needed reminders for using the sensors. **Trial outcomes** One user appreciated viewing the long-term trends	**Pilot outcomes** Users reported difficulty in configuring the system to give permission for data sharing between devices and Mashups server. Also needed help from research team to perform authentications, especially for Withings scale. Several users said when the battery died/phone restarted they'd forgotten how to get services up and running again. Not all participants noticed icon in notification tray which showed that data was being logged, meaning there were gaps in data

		liked the “objective data. Simple forms of data representation were valuable and often more appreciated than detailed graphs. Because of the lack of data provided, specific correlations for users couldn’t always be calculated. Therefore at times the system provided contradictory information. Several of our users noticed contradicting feed items over time and it led some not to trust the system and the observations in their feeds. Having more specific details than the “vague” correlations was a		which kept her coming back week after week to learn more about herself.	when service was inactive. Users did not need graphs and became confused when viewing them. Several users asked what it meant that two items were correlated - therefore users were unable to understand the coupling of the quality and accuracy of the correlations/deviations with how much data they provided. (Bentley et al., 2013; Tollmar et al., 2012) **Trial outcomes** Users reported liking the reminders as they recognized how easy it would be to forget to log. One user reported how “quick” and “easier to stick with” Health Mashups was for entering data compared to previous experience.

		common theme from our users. We had hoped that the graphs would have provided this additional data, but few users explored them in detail when encountering a feed item that they did not understand. Contradictory data was a problem for some users which lead to reduced trust in how well and reliably the system could interpret the health data. **Trial outcomes** Users described app as useful on a daily basis and take inventory and think about behaviours they could change. However others weren't convinced the data was useful			(Bentley et al., 2013)

		for making changes. This raised a tension that some participants discussed between telling people what they already know and think is “obvious” versus providing observations that are educating and perhaps a bit more prescriptive.(Bentley et al., 2013)			
iN Touch (Not available for download) PA+	NA	Technology acceptance was measured by usefulness: Mean rating for usefulness 3.50/5, SD=1.18. Participants reported the app was an effective way to see patterns about their health behaviours and make health changes.	NA	App described as ‘compelling'. Participants enjoyed the ability to take pictures and write notes in a journal and add other trackers from TheCarrot.com.	**Design test outcomes** All YAB testers (Youth Advisory Board) were able to navigate the application, enter and retrieve data immediately after the brief overview by the research assistant. They found the iPod touch to be intuitive and the application easy to use and understandable. **Implantation**

		Participants expressed that the intervention fit their needs and they felt they were well supported during the study. A few noted that they tended to do the same activities and eat the same foods so tracking repeatedly seemed redundant. (Kim et al., 2015)			**outcomes** Mean score for ease of use 3.85/5 SD=1.27. Technology acceptance as measured by usefulness, ease of use and use appears to have been quite good. Participants agreed that it was easy to learn and use throughout the program. They indicated that the first meeting in which they were given a brief training on the application was useful and allowed them to adopt the technology quickly. Most of the youth found the application easy to use and the iPod convenient and portable. Almost all participants reported that the exercise, mood and socialising trackers were very easy to use and took little time. The food

					tracker was the most difficult to use. Participants felt it was difficult to find the exact food they were looking for and they did not enjoy having to enter each ingredient. Some participants reported that they used the journal feature or picture feature for food as these features required fewer steps. Participants found themselves at times exiting out of the application before they hit the save button and not realizing the data were erased. (Kim et al., 2015)
Motimate (Not available for download) PA+	NA	NA	NA	NA	NA
Ngala (Not	NA	NA	NA	NA	NA

available for download) PA+					
SIGMA (Not available for download) PA+	The palette of colours was appreciated as well as the touch interactive map embedded in the SIGMAe. However, users argued for better integration of the map within the story line of the game. The 2D avatars and 2D environments that we used within SIGMAe were highly appreciated. Users requested more brightness adjustments to the 2D environments and a better randomization of the food related scenarios. Users	Perceived utility and degree of learning regarding how much the participants learned in terms of managing sabotaging thoughts and unhealthy behaviours with regard to food temptations, the majority of the participants (90.5%) estimated that they learned “a great deal”, while 90.4% of the participants reported that the SIGMA application is “useful” and “very useful” in resisting food temptations. Moreover, 76.2% of the participants reported that they would use the	NA	NA	**Pilot outcomes** Analysis revealed a mean score of 84.40 (SD = 10.72; range 0 - 100) for the overall usability of the application, based on SUS (System Usability Scale). Some of the highest ranked items were “I would imagine that most people would learn to use the SIGMA application very quickly” and “I thought the SIGMA application was easy to use”. The majority of the participants were very satisfied (scores of 5 or 6 out of 6) with the application’s user interface in terms of figuring out how to play (85.7%), understanding the instructions (95.2%),

	wanted a more symmetrically and evenly distributed text within the psycho-education module. Participants stated that in order to promote usability, information must be clearly provided with a consistent design throughout the app.	application for weight gain prevention, while 81% reported that they would use the application for attaining weight loss. (Podina et al., 2018)			ease of knowing for whether they were doing good or not (95.2%). Users wanted a simple, fast app with a minimal amount of effort invested in searching throughout the menu and within the app. Hence, features that were automated, such as the pedometer, the progress bar, the healthy tips that we provided, and the automated feedback delivered after playing the game were appreciated. Several participants reported becoming quickly frustrated when a feature was not intuitive enough to use, such as the back button in the SIGMAe game. Some found the 10-minute daily request to access and use the app

					as too long or too frequent. Others found it cumbersome to use the self-monitoring feature of the app, finding it time-consuming and difficult to identify tempting situations or to identify and enter the thoughts that were triggered by those situations. Many participants required reminders to prompt use. A few required less scrolling to find information in the self-monitoring feature and more intuitive back and forward buttons to navigate within SIGMAe and SIGMAi. (Podina et al., 2018)
One You (Downloaded) PA	NA	NA	NA	NA	System usability scale score 71/100. (Bondaronek et al., 2018)

Runkeeper (Downloaded) PA	Reported as pros: Attractive, user-friendly design.	Reported as cons: Not validated. It does not use the smartphone accelerometer.(Martinez-Nicolas et al., 2017)	Preliminary results showed that certain app elements such as goal setting, and number of likes positively influence users' physical activities.	Reported as pros: Gamification using social networks and challenges. (Martinez-Nicolas et al., 2017)	Reported as pros: User-friendly design. Available in 13 languages.(Martinez-Nicolas et al., 2017) Runkeeper was tested to see whether it met 18 ergonomic criteria across 15 tasks. Summarizing, RunKeeper satisfied 70.37 % of the criteria, 9.26 % of its tasks were not satisfied. RunKeeper could be considered easier to use for initial users. Also, RunKeeper had 20.37 % of its tasks not available or not applicable to evaluate. (Klock and Gasparini, 2015)
Runtastic HR (Downloaded) PA+	NA	Correlated well with ECG for regular rhythms but worse for those with irregular rhythms or	NA	NA	NA

		tachycardia. App was worse in both respects that Instant HR and Cardiio apps. (Pipitprapat et al., 2018) Runtastic Heart Rate uses contact method (phone on finger) to measure HR and seems to produce results in a quite acceptable range and did not produce any reading until a valid human body part was detected. Runtastic outperformed most other tested apps on mean square errors for accuracy of recordings taken during the day and after waking up, only one outperformed Runtastic on waking			

		up measurement. Application’s accuracy is platform-independent. (Pipitprapat et al., 2018) Although the apps tested (including Runtastic HR) had greater correlations at certain time points, these results were not consistent or significant enough to suggest that the apps are accurate or valid compared to an ECG. The Runtastic Heart Rate app had inconsistent correlations. (Papon et al., 2015) Runtastic had relatively weak correlations to the standard ECG and Polar® T31			

		uncoded heart rate monitor. Finger positions, participant movement, cold or sweaty fingers and phone case may be problematic. (Bouts et al., n.d.)			
Runtastic (Downloaded) PA+	NA	Findings suggest that Runtastic grossly underestimates the steps walked and distance covered. If Runtastic is being used, it should be kept in mind that it may underestimate by approximately 45%. (Poojary et al., 2018) The Runtastic app was the least accurate (Towler et al., 2018)	NA	Results indicated that achieving goals and being more motivated by enjoyment are the most important gratifications predicting engagement in a physical activity app; companionship is much less important. (Klenk et al., 2017)	Runtastic was tested to see whether it met 18 ergonomic criteria across 15 tasks. Data were only reported for the following tasks. Summary: Runtastic satisfied 70.74 % of the criteria and had 15.56 % of their tasks not satisfied. Runtastic had 13.7 % of its tasks not available or not applicable to evaluate. Runtastic could be considered more complete for trivial tasks than the others because of this. (Klock and Gasparini, 2015) System usability scale

					score 55/100 SUS. (Bondaronek et al., 2018)
Runtastic PRO (Downloaded) PA+	NA	Reported as pros: Reliability (Antón and Rodríguez, 2016)	NA	NA	Reported as pros: Easy to use, intuitive interface Reported as cons: Google play three-dimensional views fail frequently. In the top-line menu bar, only three of the four options are displayed at any given time, so you have to keep swiping to pull into view whichever display is left out at that moment (Antón and Rodríguez, 2016) System usability scale score 95/100 SUS (Bondaronek et al., 2018)
Fitbit (Downloaded) PA+	Design described as elegant, sleek, looks nice and inviting. App provided a user interface platform to	The range in correlation coefficients between nutrient intake estimates from NDSR was as follows; Fitbit,	The men liked comparing their physical activity anonymously with other participants via the pseudonym identifier on the	User experience factors, such as the design aspects of the interface and its usability, were reported as important aspects of	App combined with wearable assessed on set of tasks. 1) Time on Task: It took participants more time overall to complete the

	visualise data and review progress (Tong et al., 2018)	r=0·77–0·94. The mean nutrient intake calculation for dietary fibre was significantly lower than from NDSR (P=0·04). Energy correlation to NDSR: 0.92, carbs:0.94, protein: 0.88, fat:0.84, sodium:0.77, fibre: 0.92. Fitbit had 6 free nutrients and 10 when paid for. Source of nutrients was not disclosed and users were unable to add new foods and nutrients to database. With respect to the extent to which foods and food amounts entered into each app matched those in the NDSR foods reports, most foods	companion app: *"I think the motivation of just bein [sic] a part of a group makes you wanna [sic] do healthier things."* (Eisenhauer et al., 2017)	engagement and continued use: *"I liked the Fitbit app better—the design is certainly more elegant."* (Female, 26) (Tong et al., 2018)	tasks while using the Jawbone UP (M = 3.67 SD = 3.15) than the Fitbit Flex (M = 2.13; SD = 2.25), t(36) = 2.36, p = .022. 2) 2) Number of Steps: It took participants the same number of steps overall to complete the tasks while using the Jawbone UP (M =9.2; SD = 6.8) and the Fitbit Flex (M = 4.2; SD = 1.6), t(36) = 1.59, p = .187. 3) Errors: Participants experienced more errors with the Jawbone tasks (M=1.5, SD=1.6) than with the Fitbit

		were classified as close matches (78–83%) for each app (of which Fitbit was one). For all apps, mismatches due to poorly matched food amount (13–17 %) occurred more frequently than mismatches due to poorly matched food description (4–6 %). Poor matches (neither food amount or food description matched) were infrequent (Griffiths et al., 2018) The food log in Fitbit was considered helpful to learn eating habits. Using a 5 point Likert scale, it scored a mean of 3.45, 0.69 SD, and 55% (n=6/11) agreed it			Flex (M = 0.4, SD = 0.7), t(36) = 3.42, p = 0.002. 4) Observations: Jawbone UP participants struggled more than Fitbit Flex participants to log diet. Nine out of fourteen of the participants had experience with some kind of health logging app, possibly making this task seem more intuitive. However, a Fitbit participant realized that there was not an option to input decimal amounts (e.g. 10.5 ounces water), only whole numbers. 5) Alarm: While the

		was helpful. Fitbit app also considered helpful for learning about activity levels. Using a 5 point Likert scale, it scored a mean of 3.25, 0.62 SD, and 33% (n=4/12) agreed it was helpful. (Eisenhauer et al., 2017)			Fitbit Flex alarm function was also under the accounts tab, which could have presented confusion, like the goals did, participants did not have trouble with this task because they saw the function when setting the goals, and knew exactly where to go. 6) Workout: This task was more straightforward for those with the Fitbit. A Jawbone participant stated *“I would go to the right menu because it looks like where you input data”.* The Fitbit participants

					easily went to the “active minutes” section and searched for their exercise and selected the amount. Issues encountered during the usability test as well as participant statements tended to focus on interactions with the app rather than the band. (Altenhoff et al., 2015) Food log scored as easy to use, with a mean of 3.00, 0.89 SD while 25% (n=3/11) agreed it was helpful. App scored as easy to use - 3.58 mean, 0.67 SD, 67% (n=8/12) agree its helpful. Smart phone users liked the convenience of logging real time. The men desired a “simpler” option for tracking their

					food intake, as there were no options for home-made foods on the food log menu, mainly restaurant options. Only one man reported accessing the Companion app to log his food intake after 6 weeks. Time (peak planting season [users were farmers]) and difficulty with locating exact food-portion sizes were reasons the men cited on the survey for discontinuing logging. (Eisenhauer et al., 2017) User experience factors, such as the design aspects of the interface and its usability, were reported as important aspects of engagement and continued use. Many participants found that using the wireless tracker and scale in

					combination with a mobile app offered many advantages. Specifically, wireless devices provided an automatic way for users to collect and self-monitor personal measurements, and their integration with the mobile app provided a user interface platform for participants to visualize those data and to review progress. (Tong et al., 2018) System usability scale score 66/100. (Bondaronek et al., 2018)
Keep (Downloaded) PA+	NA	NA	NA	NA	NA
Clue (Downloaded) PA+	Positive subjective presentation score (>/=3 out of 5)	Clue included: 2/3 features indicating comprehensiveness (conception and contraception info),	NA	NA	Ease of navigation scored as being average or above average. (Moglia et al., 2016)

		it cited literature to support its content, had medical disclaimer and included health education (Moglia et al., 2016) Clue scored 0 - no criteria were met to suggest that Clue was able to accurately determine a fertility window. Calendar-based apps (including Clue) misstated the day of ovulation if the cycle length differed from the estimated cycle length in the current cycle. Some apps did not try to change the predicted fertility dates, or didn't change them enough in the current cycle (including Clue)			

		(Freis et al., 2018)			
Samsung Health (Downloaded) PA+	NA	Results indicated that the validity of Samsung Health varied depending on the smartphone model, its body location, and the type of gait (walking and running). Samsung Health showed acceptable validity when the phone was located on the hand (Bias = −8.3%; RMSE = 5.6), and especially on the arm (Bias = −7.2%; RMSE = 4.9) while running, and when the phone was located on the arm (Bias = −7.5%; RMSE = 5.4), and especially on the waist (Bias = 5.4%; RMSE = 3.7) while walking. Samsung Health only showed	NA	NA	NA

		good validity when the phone was located on the arm (Bias = 2.9%; RMSE = 3.6), and especially on the hand (Bias = 0.5%; RMSE = 2.5) while running. This application showed unacceptable validity in the remaining options. (Beltrán-Carrillo et al., 2019) The difference in estimation of energy and saturated fat intake between Dietplan6 (validation comparison) and the diet apps (of which Samsung Health was one) was not significant. Samsung Health significantly underestimated calcium, iron, and			

		vitamin C compared with Dietplan6, although there was no significant difference for vitamin A. The correlation coefficient was r=–.12 for iron (Samsung Health vs Dietplan6). Samsung Health had the greatest variation of correlation, with energy at r=.79. Estimates of micronutrient intake (calcium, iron, vitamin A, and vitamin C) by 2 apps (including Samsung Health) were inconsistent and less reliable. (Fallaize et al., 2019)			
Up app (Available	NA	Step goal was considered useful	NA	Tips and advice were scored highly	Despite technical issues such as broken monitors

version not functioning) PA+		on a 5 points Likert scale, scoring a mean of 4.95 (SD 0.23). Users would continue using step goal, scoring a mean of 4.84 (SD 0.50). Information was considered credible, scoring a mean of 4.63 (SD 0.83). Information was considered relevant, scoring a mean of 4.63 (SD 0.60). (Lyons et al., 2017)		as being perceived as specific to user: Mean 4.37 (SD 1.01). (Lyons et al., 2017)	(not syncing, not powering on, buttons falling off), participants reported that the monitor, tablet, and app were user-friendly. App was also scored as user-friendly with a mean of 4.68 (SD 0.58). (Lyons et al., 2017) App combined with wearable assessed on set of tasks. 1) Time on Task: It took participants more time overall to complete the tasks while using the Jawbone UP (M = 3.67 SD = 3.15) than the Fitbit Flex (M = 2.13; SD = 2.25), $t(36) = 2.36$, $p = .022$. 2) 2) Number of Steps: It took participants the same number of

					steps overall to complete the tasks while using the Jawbone UP (M =9.2; SD = 6.8) and the Fitbit Flex (M = 4.2; SD = 1.6), $t(36)$ = 1.59, p = .187. 3) Errors: Participants experienced more errors with the Jawbone tasks (M=1.5, SD=1.6) than with the Fitbit Flex (M = 0.4, SD = 0.7), $t(36)$ = 3.42, p = 0.002. 4) Observations: Jawbone UP participants struggled more than Fitbit Flex participants to log diet. Nine out of fourteen of the participants had

					experience with some kind of health logging app, possibly making this task seem more intuitive. However, a Fitbit participant realized that there was not an option to input decimal amounts (e.g. 10.5 ounces water), only whole numbers. 5) Alarm: While the Fitbit Flex alarm function was also under the accounts tab, which could have presented confusion like the goals, participants did not have trouble with this task because they

					saw the function when setting the goals, and knew exactly where to go. 6) Workout: This task was more straightforward for those with the Fitbit. A Jawbone participant stated “I would go to the right menu because it looks like where you input data”. The Fitbit participants easily went to the “active minutes” section and searched for their exercise and selected the amount. Issues encountered during the usability test as well as participant statements tended to focus on interactions

					with the app rather than the band. (Altenhoff et al., 2015)
TOTAL	**4/22 apps reported feedback on aesthetics**	**4/22 apps reported positive feedback on credibility** **3/22 apps reported negative feedback on credibility** **10/22 apps reported mixed feedback on credibility**	**3/22 apps reported feedback on effectiveness of app or app components**	**10/22 apps received feedback on engagement** **5/22 received positive feedback on engagement** **4/22 received negative feedback on engagement** **1/22 received mixed feedback on engagement**	**14/22 apps reported feedback on functionality**

2D = two dimensional, ECG = Electrocardiogram, HR = Heart Rate, M = Mean, NA = Not applicable, NDSR = Nutrition Data System for Research, PA = Physical Activity/Physical activity targeted by app, PA+ = Physical activity and other behaviours targeted by app, RMSE = Root Mean Square Error, SD = Standard Deviation, SUS = System Usability Scale

Appendix 19 Strengths and weaknesses of physical activity measures used in apps

App	Evidence provided by author/developers for PA measure	Strengths of measure	Weaknesses of measure
Literature-based apps			
Haptivity (Not available for download) PA	Justification for not providing information on PA levels: implied that providing quantification of PA appeals to those who get pleasure from quantification and therefore focusing on the pleasure of the activity itself might be good for other, sedentary, individuals (Forster et al., 2017)	NA	Users stated that they wanted information on their PA levels (Forster et al., 2017)
MAPS (1-4) (Not available for download) PA	*"Tracking is common in effective digital health interventions (Glynn et al 2014, Hurling et al 2007, Schoeppe et al 2016)"* (Fanning et al., 2017)	In-app tracking was a favourite feature of users (Fanning et al., 2017) Self-monitoring strengths.	Couldn't do retrospective tracking, which users wanted (Fanning et al., 2017) Lack of frequent activities menu too - make it easier (Fanning et al., 2017) Wanted integration with activity monitors (Fanning et al., 2017) Users didn't like having to manually track data every day (Fanning et

			al., 2017) Self-monitoring weaknesses. - Data accuracy - Understanding type, intensity of activities (e.g. aerobic versus non-aerobic). Although orientation did try to explain (Fanning et al., 2017)
Unnamed app (Not available for download) PA+	*“The accelerometer has been employed in several applications to detect human movement detection, to characterize physical activity (Shin et al., 2010) recognising if the user is walking, going upstairs or running (Brezmes et al., 2009).”* *“Reducing the number of frequency bins which must be searched to acquire the GPS signal reduces the Time To First Fix (Jarvinen et al., 2002)”* (Fernández et al., 2013)	*“The main advantage of these sensors is that they don’t require an external communication channel, such as Bluetooth or WiFi improves the battery consumption and the facility to access the data provided by the sensor.”* *“In our application we have developed filters to reduce noise; Butterworth second order high-pass filter is implemented.”* *“The accelerometer gives three axis forces applied on the smartphone, in the implemented application the contribution of earth gravity have been removed.”*	*“The main limitation of GPS_provider is that it only can work in outdoor environment, because in indoors GPS signals don't reach. Moreover this provider consumes more battery than the others. The main disadvantage of Google Maps that it has a price policy of use otherwise OSM is totally free of use*.”(Fernández et al., 2013) Objective monitoring weaknesses.

		"Use of multiple location providers for different GPS scenarios. GPS_provider determines location using satellites. Depending on conditions, this provider may take a while to return a location fix, this time is called Time to First Fix (TTFF.) This time can be cut down substantially by employing Assisted GPS (A-GPS). A-GPS uses the mobile network to transmit the precise GPS satellite orbit and clock information to a mobile device. Then the device will know the approximate location of the GPS satellites in its line of sight. Reducing the number of frequency bins which must be searched to acquire the signal reduces the TTFF (Jarvinen et al., 2002). GPS_provider provides the most accurate location data and it is unique that has ability to determine altitude information. Therefore, we need use GPS_provider to track people exercise properly." *"In order to improve the accuracy of this information, user can pause and resume the tracking, this is essential*	

		to avoid wrong results." (Fernández et al., 2013) Objective monitoring strengths.	
ATHENA apps (Not available for download) PA+	*"The activity recognition module enabled position-free recognition and was able to recognise activities wherever the smartphone was attached to the body. The evaluations showed that the system works well in real-world environments with an accuracy of 92.43%."* (Fahim et al., 2014) No information about step counter or wearable.	Wearable specifically targets posture and gesture activities in sports (Fahim et al., 2014) *"The prompt labelling module is better way to ensure 'ground truth' of the data being collected by the activity recognition module than other methods listed in the paper as it is more representative of a real-world method of data validation, less laborious and time-consuming, more feasible for long-term real-world situations, better for data privacy and scalability, more likely to ensure correct and entered data as prompts to label occur when user is standing, not in the middle of an activity, doesn't require so much battery as voice recognition and avoids voice recognition inaccuracies. It also has more integrity as labels are automatically saved after a respective activity. Validation via labelling shows the trustworthiness of the activity*	More frequent recording of GPS coordinates gives more location accuracy but on the other hand puts a bad impact on battery life of recording device (Saleem et al., 2012) Unclear if the activity recognition module prompts to label are being used still in ATHENA platform. Self-report measure weaknesses. Objective measure weaknesses as well as wearable weaknesses.

		recognition module as well as a fully annotated activity data set." (Cleland et al., 2013) Objective measure strengths.	
Health Mashups (Not available for download) PA+	Not reported	*"Devices such as the Fitbit and Nike+ sensors have allowed people to examine their physical activity at a great level of detail. Similar devices such as Philips Direct Life provide easy ways to understand daily activity levels and provide simple suggestions on ways to be more active throughout the day. Internet-connected scales (e.g., the popular Withings model) allow people to easily keep track of their weight & changes over time without the need for manual log-keeping."* (Tollmar et al., 2012) Fitbit - easily available for general population, automatically uploaded step-count, allows examination of PA in detail. Five star scale - authors reported it as low-effort (Bentley et al., 2013)	*"While some of these services, such as Fitbit, allow users to import data from multiple sensors (e.g., Fitbit and Withings) into a single account, these commercial services currently do not provide any graphs, insights, or suggestions to users based on the combination of different wellbeing data feeds. Each sensor is devoted to its own space in the interface. For example, graphs on the Fitbit website show information regarding the number of steps the user has walked in one box and a graph of the user's weight in another with no way to directly compare them or to easily discern patterns in the data over time."* (Tollmar et al., 2012) Fitbit - pilot users did not wear it all day or just took it out to use when exercising, problems authenticating and setting up the measure to link

		Self-report measure strengths. Objective measure strengths.	with the Mashup server. Five star scale - not often used, not clear where it comes from (Bentley et al., 2013) Self-report measure weaknesses. Objective measure weaknesses.
iN Touch (Not available for download) PA+	Not reported	Youth Advisory Board wanted minimal data entry, which app accommodates: short to complete and minimal taps. Self-report measure strengths.	Unclear how easy it is to use without WiFi. Self-report measure weaknesses.
MotiMate (Not available for download) PA+	Not reported	Seems simple and brief and fairly intuitive. Can enter as many entries as user likes and not restricted by what a monitor would/can pick up only. Self-report measure strengths.	Not sure it explains intensity though asks user to make a judgement on that as well as duration of exercise. Not clear where the question comes from at all or if it's been tested. Self-report measure weaknesses.
Ngala (Not available for download) PA+	Only evidence for it came from developmental qualitative work that suggested that potential users wanted information relevant to their	Unclear if its objective or subjective	Unclear if it's objective or subjective. Complete lack of details.

	individual issues, including through user self-assessment tools and ongoing tracking of their progress (Hearn et al., 2014)		Physical activity self-assessment over a 1 year period was low compared to other forms of self-assessment suggesting some potential issues either of engagement or functionality.
SIGMA (Not available for download) PA+	*"The decision to incorporate a pedometer was informed by the fact that it has been reliably associated with significant increases in physical activity and significant decreases in BMI [43]."* (Podina et al., 2017)	Automatic collection appreciated by users. Objective measure strengths.	Objective measure weaknesses.
Descriptives:	**9/12 apps reported justification for their method of PA measurements/capture or lack of measure**		
Public apps			
One You (Downloaded) PA	NA	NA	Did not activity measure PA.
Runkeeper (Downloaded) PA	Not reported	*"GPS and stopwatch mode for recording is good, objective. App can record different activities*." (Martinez-Nicolas et al., 2017) Self-report measure strengths. Objective measure strengths.	*"Not validated, doesn't use smartphone accelerometer. Can only record certain activities and certain details of them. GPS signal can be lost*." (Martinez-Nicolas et al., 2017) Self-report measure weaknesses.

			Objective measure weaknesses.
Runtastic HR (Downloaded) PA+	NA	NA	PA not reported at all.
Runtastic (Downloaded) PA+	Not reported	Self-report measure strengths. Objective measure strengths.	Inability to report activities not listed in app. Records activity based on duration irrespective of movement resulting in questionable accuracy of data. GPS signal can be lost. *"Runtastic underestimates step count".* (Poojary et al., 2018) Self-report measure weaknesses. Objective measure weaknesses.
Runtastic PRO (Downloaded) PA+	Not reported	Self-report measure strengths. Objective measure strengths.	Inability to report activities not listed in app. Records activity based on duration irrespective of movement resulting in questionable accuracy of data. GPS signal can be lost. Self-report measure weaknesses. Objective measure weaknesses.
Fitbit (Downloaded)	Not reported	Coach also logs activities completed without including those that have	Steps don't register from mobile tracking immediately and are also

PA+		been skipped so accuracy is better. Self-report measure strengths. Objective measure strengths.	included from manually logged activities that may be misleading. Self-monitoring only allows addition of certain activities. Coach will think you've completed activities if you just let the app run. GPS signal can be lost. Self-report measure weaknesses. Objective measure weaknesses.
Keep (Downloaded) PA+	Not reported	Automatic step counter and training plan completion appears effortless. Training plan activity capture only captures those activities that run to the end. Self-report measure strengths. Objective measure strengths.	Inability to report activities not entered in list. Step counter doesn’t necessarily work immediately to show results. Activities and steps don’t appear in training insights straight away. Also have to ensure that phone is set up to allow step counts to be extracted to app. Training plan will report you've done activity if you just leave it running - doesn’t guarantee it’s been done by user. Self-report measure strengths. Objective measure weaknesses.
Clue	Not reported	Simple and quick and easy to do.	Only allows tracking of a small

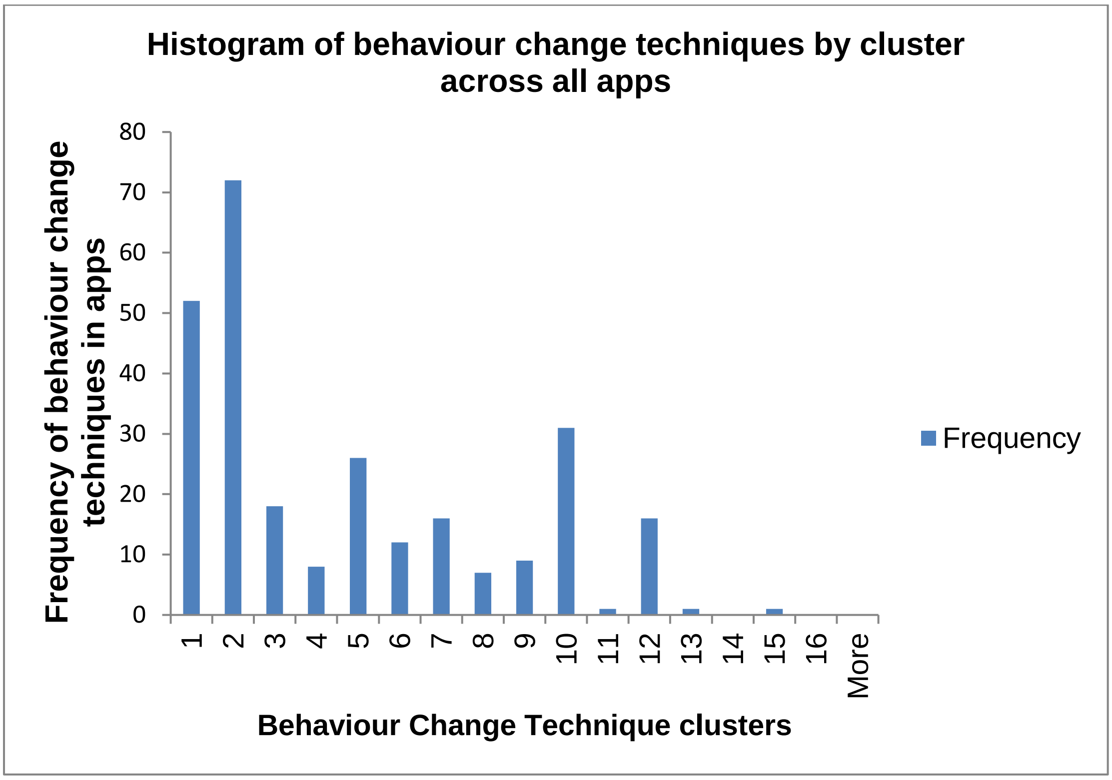
(Downloaded) PA+		Self-report measure strengths.	subset of activities and no characteristics of those activities such as distance, duration, intensity. Self-report measure weaknesses..
Samsung Health (Downloaded) PA+	App gives potential reasons for step count errors - including suggestion that phone count might differ to a wearable.	App provided exercise programme tracking doesn't register completion if you skip through the exercises. Self-report measure strengths. Objective measure strengths.	Lots of activities but no 'other' option available if you activity is not listed. *"The validity of Samsung Health varied depending on the smartphone model, its body location, and the type of gait (walking and running)."* (Beltrán-Carrillo et al., 2019)
Up app (Available version not functioning) PA+	*"Accelerometer-based tracking devices and apps, especially those that facilitate the use of theory-driven behaviour change strategies, have the potential to impact physical activity–related behavior.14* *Researchers have recognised that technology, including wearable fitness tracking devices, such as the Jawbone UpBand, have the potential to be used as a mechanism to motivate individuals'*	Automatic and passive data collection via the UpBand (Harris et al., 2018; Melton et al., 2016) Objective measure strengths.	*"Participants were required to take on additional responsibilities by using an accelerometer and app, which could have been perceived as more of an obligation or punitive, thus inhibiting the perception of autonomy within the intervention.* *Attention is warranted regarding the validity associated with changes in physical activity, as well as psychosocial constructs associated with physical activity*

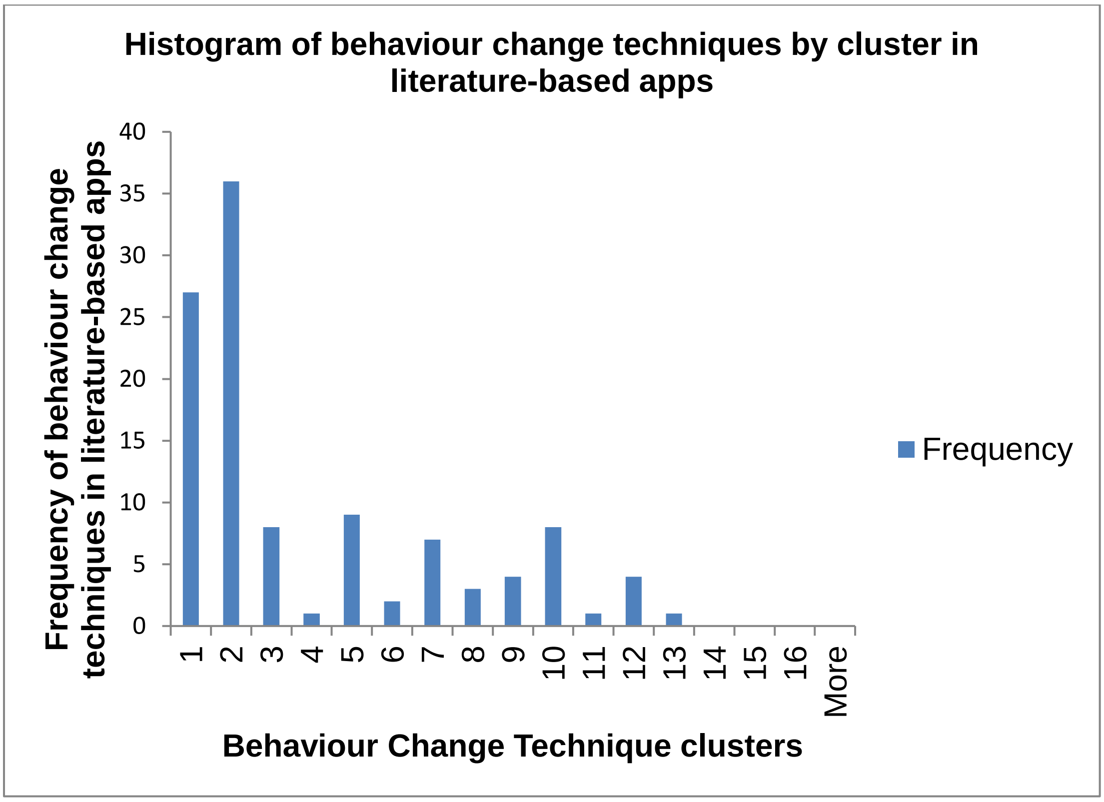
	physical activity behaviors.15 A recent systematic review of electronic activity monitors noted a paucity of research regarding their efficacy. 40" (Harris et al., 2018) *"The UP Band effectiveness has a high degree of accuracy with step counts and sleep tracking and is comparable with other commercial products (Rosenberger, Buman, Haskell, McConnell, & Carstensen, 2016)."* (Melton et al., 2016)		*adoption and maintenance from use of these devices."* (Harris et al., 2018) Requires wearable, not just app. So users must remember and charge it. Objective measure weaknesses.
Descriptives	**2/10 apps reported justification for their method of PA measurements/capture or lack of measure**		
Total	**11/22 apps reported justification for their method of PA measurements/capture or lack of measure**		

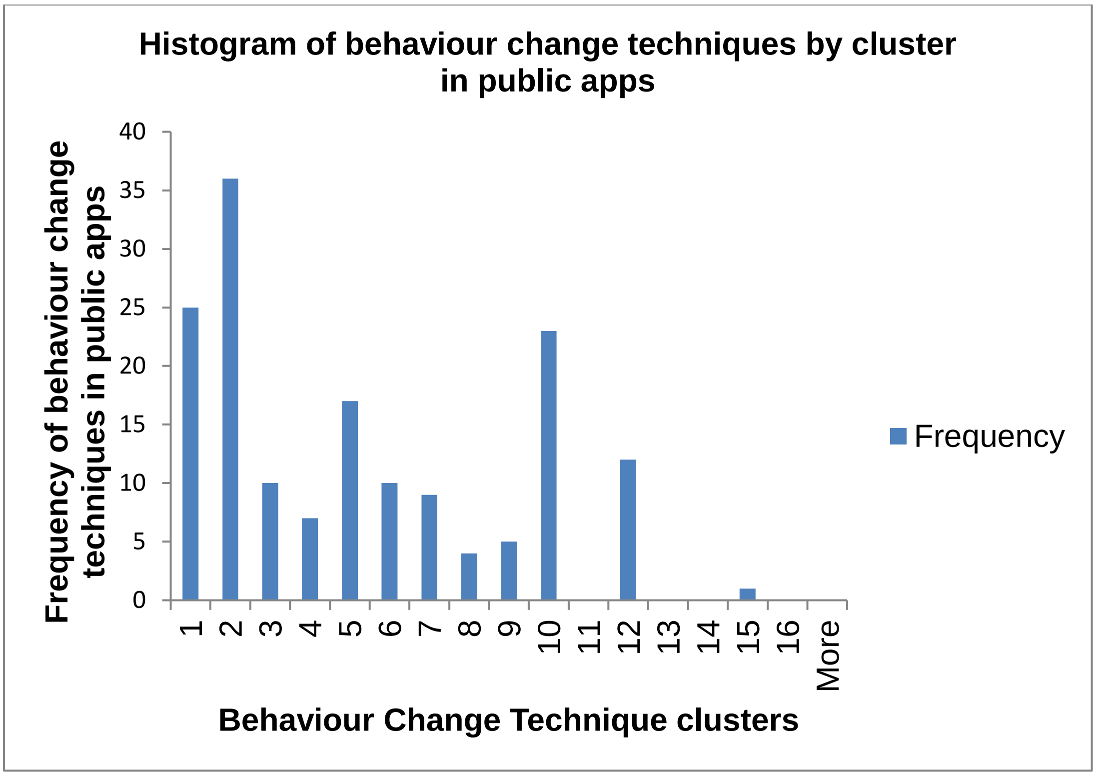
Italicised text is quoted verbatim from paper.

BMI = Body Mass Index, GPS = Global Positioning System, NA = Not applicable, OSM = Open Street Map, PA = Physical Activity/Physical Activity targeted by app. PA+ Physical activity and other behaviours targeted by app, TTFF = Time To First Fix,

Appendix 20 Histograms of behaviour change technique clusters in apps







Clusters:

1. Goals and planning
2. Feedback and monitoring
3. Social support
4. Shaping knowledge
5. Natural consequences
6. Comparison of behaviour
7. Associations
8. Repetition and substitution
9. Comparison of outcomes
10. Reward and threat
11. Regulation
12. Antecedents
13. Identity
14. Scheduled consequences
15. Self-belief
16. Covert learning

Appendix 21 Analysis of positive and negative user reviews

App	Negative reviews (n, %)	Positive reviews (n, %)	Both (n, %)
One You (Downloaded) PA	3 (6%)	39 (78%)	8 (16%)
Runkeeper (Downloaded) PA	11 (22%)	33 (66%)	6 (12%)
Runtastic HR (Downloaded) PA+	9 (20%)	19 (42.22%)	17 (37.78%)
Runtastic (Downloaded) PA+	8 (16%)	41 (82%)	1 (2%)
Runtastic PRO (Downloaded) PA+	7 (14%)	35 (70%)	8 (16%)
Fitbit (Downloaded) PA+	38 (76%)	9 (18%)	3 (6%)
Keep (Downloaded) PA+	3 (6%)	40 (80%)	7 (14%)
Clue (Downloaded) PA+	10 (20%)	25 (50%)	15 (30%)
Samsung Health (Downloaded) PA+	8 (16%)	36 (72%)	6 (12%)
Mean	**10.78**	**30.78**	**7.89**
%	**21.8**	**62.25**	**15.96**
Total	**97**	**277**	**71**

Appendix 22 Themes, subthemes and exemplar quotes from user reviews of public apps

Theme/Subtheme	Example quote
1. Acceptability	
1a. Recommendable	**+ve -** i love this app its perfect its really good i really recomend it for girls like thid is the best app for period tracking on earth i luv it u really should consider downloading it **-ve -** Options are too limited and I can't even switch it to pregnancy mode. Wouldn't recommend this as the scope is too limited.
1b. Preference for or absence of specific content	**+ve –** Great app to keep a record for runners **-ve -** Way too many ads and too many "share your workout" pushes. It's ridiculous I'm getting bombarded with push notifications every 5 mins. I know it's part of marketing but damn that's just absurd.
1c. Positive experience	**+ve -** I feel great. I was insecure about going to the gym or to a fitness class, but this app makes me feel exercised and like I'm doing something with my life. I love this app . i feel so safe on this calendar track . i Feel free 😊❤ HAHAHA LOL .
1d. Cost and upselling	**+ve -** No need for a gym. I don't know why I didn't found this sooner. Literally a personal trainer in the app form, saving so much money doing everything at home. This is by far the best fitness app I've ever downloaded and would definitely keep. **-ve -** just installed and bam you got me with an upgrade screen even before i can look at what the app can do for me. custom plans, sounds interesting, bam, another upgrade screen. thanks but no thanks
1e. Compatibility with devices	**+ve –** Works well with my iphone se. **-ve -** Ok but… Am I the only person that thinks the app has missed something? Like I dunno maybe utilising apple watch! Why no apple watch integration? Seems like a massive flaw t me

1f. Resource heavy	**-ve -** The application is ok, but it using 30% of phone battery is completely unacceptable. even disabling location tracking and other unnecessary features does not solve this problem. it has to go
2. Aesthetics	**+ve -** Awesome concept, visually sleek, Overall, love the format and neutral colors as well!
3. Credibility	
3a. Accuracy/Reliability	**+ve-** I switched MapMyRun a few years ago but recently it was not accurately plotting my courses as it only appears to mark coordinates too far apart. I've tried various recommended options to no avail. when I use Runkeeper concurrently it works great while Mapmyrun continues to fail. I'm switching back to Runkeeper. **-ve-** i take the same route everyday but this app shows different distances and routes everyday.you just had one job to do and you screwed up
3b. Safe/Trustworthy	**+ve-** Very good. This is a very good app and I have used a few.. I have HCM so its good to keep a check. **-ve-** This app is not helpful at all I ended with a shoulder pain.
3c. Usefulness	**+ve-** audio alerts are useful in reviewing the activity fast and act accordingly. **-ve-** My only criticism is that the trainer doesn't mention that it's beneficial to do stretching exercises after each run.
4. Effectiveness	
4a. Motivational	**+ve -** gets my fat butt up and out excellent app. keep you motivated and urges you to go at you own pace.
4b. Physical change/effect	**+ve -** It really works! On week one I thought I'd die running for 60 seconds, now I'm running for 5 minutes and could do more! I've completed the beginners three week programme and already lost weight, with friends commenting on how I already look more toned (after only THREE weeks!!!)
5. Engagement	**+ve -** I love this because it makes me very competitive to get #1

	I love how they engage and arrange everything, wants me to go back and train everyday
5b. Tailoring/Suitability	**+ve -** Variety of exercises suitable for everyone. Easy to use app to choose the right ones for you.
6. Functionality/Usability	
6a. Ease of use	**+ve-** Just finished week 1. Really straightforward to use. **-ve -** Jus don't understand how to work this app, tried a few times now I'm giving up
6b. Functional errors/lack of errors	**+ve-** Good. No complaints here. Works well for me on my 6S **-ve –** Crashes throughout workout 😒
6c. Misleading	**-ve -** Claims to be a route planner app, but this feature is not available with the free version.. Misleading
7. Security/Privacy	**-ve -** Profile is not redactable!
8. Other	**-ve -** Wrongly subscribed. This is a very good fitness app but I accidentally subscribed/paid for a year and I can't find any contact email or anything to get in touch with the company to undo this mistake. **+ve -** oh great for kids

Spelling and grammatical errors taken verbatim from user reviews

Appendix 23 Correspondence with literature-based app author 1

Email received: 25.1.19

Hi Laura,

[redacted]

Unfortunately we haven't got any further with the app. My contract has made it difficult for me to apply for further funding as PI and it hasn't been the priority of my collaborators. I also no longer have access to the app. A developer was maintaining it, but we haven't made it available to new iOS software and so it won't work anymore without edits (and no money to pay for edits)!

[redacted]

Appendix 24 Correspondence with author of an excluded app

Email received: 13.11.18

I agree. It think it's stupid, given the amount of money and energy that has been spend.

But the company who paid for it decided to shut it down.

[redacted]

[redacted]

[redacted]

Appendix 25 Correspondence with literature-based app author 2

Email received: 15.4.19

Hi there Laura,

[redacted] Because the goal of these projects is to create knowledge around specific useful app features, we test concepts in modular [redacted] apps. This unfortunately means that [redacted] app used in that paper is not standing any longer.

[redacted]